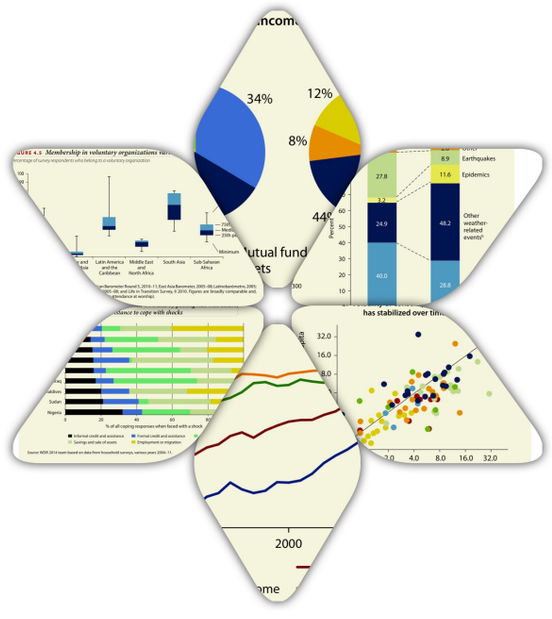
Duccio Schiavon
Luca Giuliano

WIZARD GRAFICO

Una guida alla visualizzazione grafica dei dati numerici



Data Science

Duccio Schiavon
Luca Giuliano

WIZARD GRAFICO

Una guida alla visualizzazione grafica dei dati numerici

Data Science

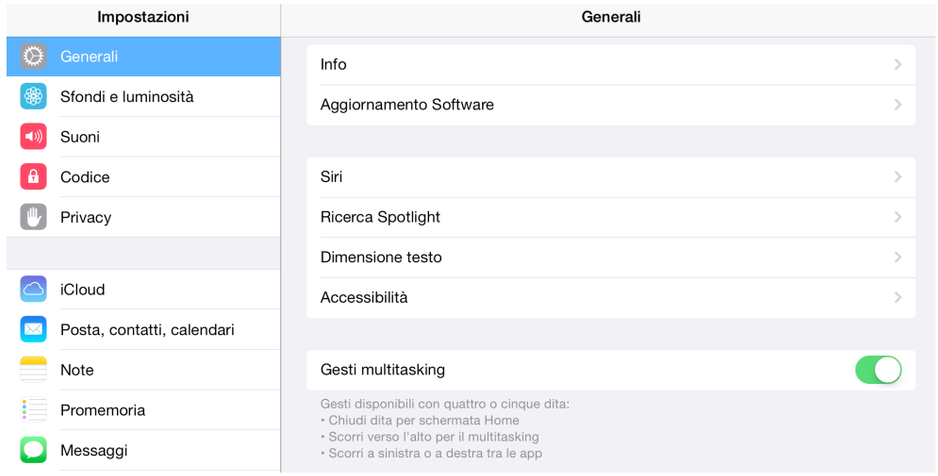
Duccio Schiavon - Luca Giuliano
Wizard grafico. Una guida alla visualizzazione grafica dei dati numerici
Roma : Dipartimento di Scienze statistiche [2014] 69 p.
ISBN 978-88-908757-1-7

Questo libro è stato realizzato con iBooks Author per la visualizzazione in Apple Mac con OS X Mavericks e iPad con iOS 5 o versioni successive. Per la segnalazione di errori e imprecisioni scrivere agli autori: Duccio Schiavon, info@stat-project.com o Luca Giuliano, luca.giuliano@uniroma1.it. Una copia statica in PDF è disponibile sul sito del dipartimento di Scienze statistiche - Sapienza Università di Roma: http://www.dss.uniroma1.it/ricerca/pubblicazioni

Istruzioni sull’attivazione dell’opzione “Gesti Multitasking” su iPad

Per una lettura ottimale di questo libro interattivo e per una piena fruizione dei link esterni ci dobbiamo prima di tutto assicurare che l’opzione Gesti multitasking per iPad sia abilitata.

Aprite **Impostazioni** e toccate **Generali**. Scorrete verso il basso per trovare “Gesti multitasking” e posizionatelo su ON.



Utilizzando quattro o cinque dita per strisciare in orizzontale (come nella figura al centro, qui sotto) si vedranno scorrere le applicazioni aperte. Provate scorrendo da destra a sinistra, dal momento che molto probabilmente in questo momento siete nell’ultima applicazione utilizzata.

Questo gesto vi permetterà di ritornare facilmente alla lettura del libro nel punto in cui vi siete interrotti dopo aver seguito un link esterno nel browser Safari.

Le altre funzioni di Gesti multitasking, spesso poco note agli utenti di iPad, sono:

- Chiudere le applicazioni e tornare alla schermata principale con un gesto di “pizzico” effettuato con quattro o cinque dita (figura a sinistra).
- Rivelare la barra delle applicazioni aperte con un colpo in su effettuato con quattro o cinque dita (figura a destra). Ripetendo il colpo verso il basso la barra delle applicazioni si chiude.

Capitolo 1

Introduzione

L’idea all’origine della realizzazione di questo Wizard Grafico è nata dalla lettura di un post pubblicato da Amit Agarwal sul suo blog *Digital Inspiration* dal titolo “How to Find the Right Chart Type for your Numeric Data”, in cui è riportato un semplice ma ingegnoso diagramma da lui ideato.

1

Cosa desideri mostrare?

Agarwal ha tracciato uno schema (fig. 1.1) che ha la funzione di suggerire il tipo di grafico più adatto sulla base dei dati a disposizione e del particolare scopo rappresentativo. Lo schema ha origine da una **domanda centrale**:

Cosa desideri mostrare?

Si tratta di un semplice interrogativo dalla risposta, tuttavia, per nulla scontata. Oltre a essere semplice, questa domanda potrebbe infatti apparire smisuratamente generica, e contemplare perciò un numero illimitato di soluzioni. Nel diagramma non è indicato il contesto analitico, il settore di applicazione, né con quali strumenti ottenere la rappresentazione desiderata: manca cioè un'ulteriore indicazione fondamentale che consenta di limitare il numero di possibili risposte. Nella nota introduttiva al post è lo stesso Agarwal a circoscrivere le possibilità di applicazione presentando il diagramma come uno strumento ideato per la rappresentazione esclusivamente di **dati numerici**, tuttavia la nota non è sufficiente da sola a chiarirne completamente il significato. Solo osservando bene tutti i suoi suggerimenti grafici ed identificandone gli elementi è possibile comprendere come interpretare correttamente il diagramma di Agarwal.

Prima di tutto si può notare come il limite al numero di possibili risposte sia senza dubbio quello definito dalla **statistica**: i grafici suggeriti dal diagramma come ideali per il particolare tipo di analisi sono tradizionalmente impiegati per la rappresentazione di misure e aggregazioni numeriche frutto di elaborazioni statistiche. Così pure la terminologia utilizzata (**distribuzione**, **composizione**, **variabile**, **differenza relativa** e **differenza assoluta**, ecc.) appartiene ad un contesto prevalentemente statistico, nonché la stessa organizzazione per schemi del percorso logico che porta al risultato finale (rappresentazione) ha in sé qualcosa di peculiarmente statistico.

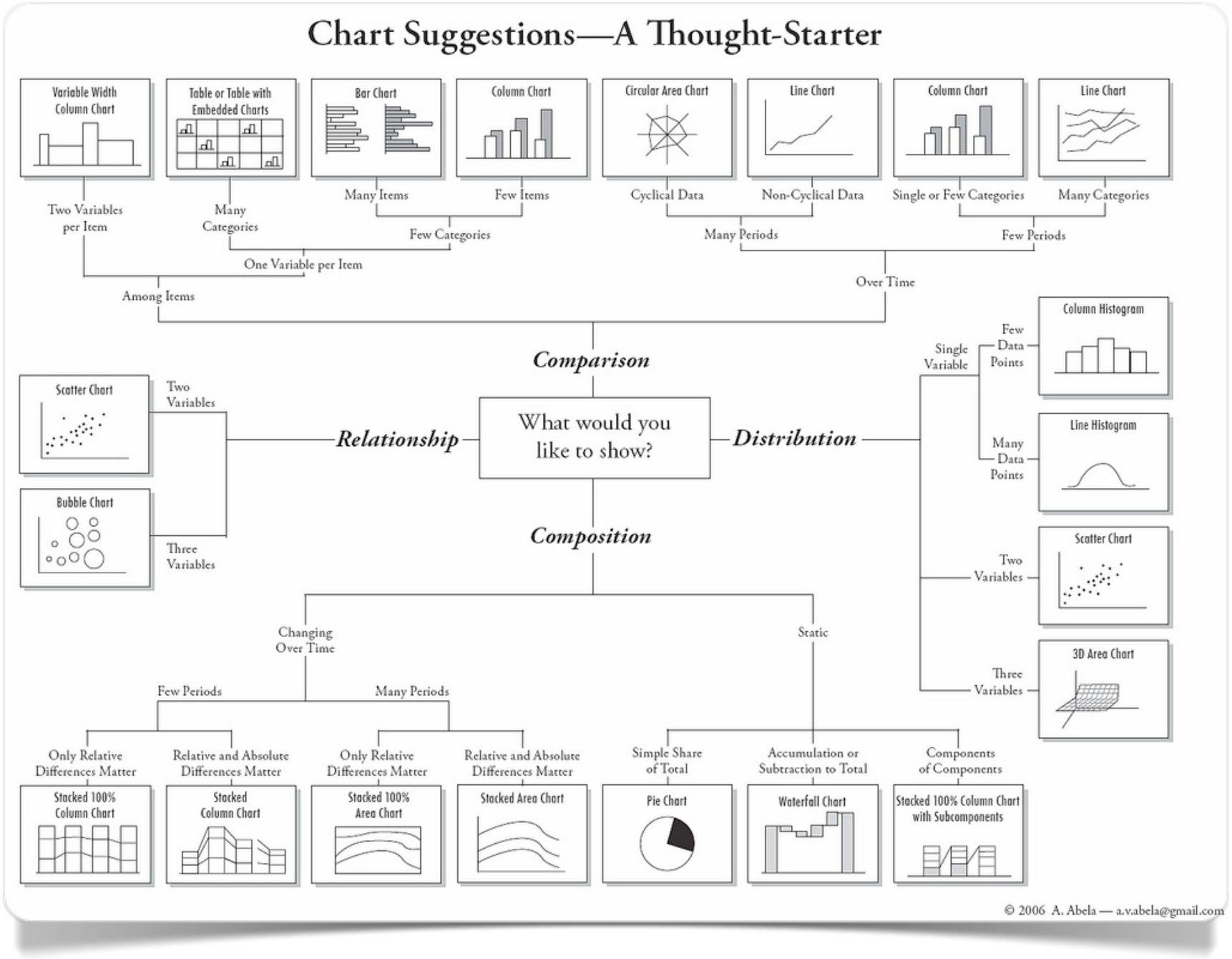


Fig. 1.1 Diagramma di Agarwal

Personalmente apprezzo molto la semplicità del diagramma. Alla domanda centrale di partenza “Cosa desideri mostrare?” segue la possibilità di scegliere tra quattro risposte (“Relazione”, “Confronto”, “Distribuzione”, “Composizione”) secondo la logica tradizionale dei diagrammi di flusso. Una volta scelta la prima risposta, vengono posti ancora altri interrogativi, da un minimo di 1 ad un massimo di 3 a seconda della prima risposta fornita.

Dopodiché, sulla base del percorso scelto all’interno del diagramma di Agarwal, all’utilizzatore viene indicato un disegno stilizzato del tipo di grafico (statistico) da utilizzare sulla base delle risposte date lungo il percorso.

Il Wizard Grafico è chiaramente ispirato al diagramma di Agarwal ed intende rappresentare un primo tentativo d’ideazione di un metodo applicabile ai più moderni strumenti di consultazione dell’informazione (*tablet*, *mobile device*, ecc.) e allo stesso tempo conduca l’utilizzatore attraverso un facile accesso ad argomenti decisamente tecnici. Naturalmente gran parte dell’efficacia del diagramma di Agarwal risiede nell’estrema capacità di sintesi: caratteristica che accomuna molti dei prodotti della moderna arte dell’**infografica**.

Come tutte le raffigurazioni pensate per il web, il diagramma di Agarwal non può tuttavia riuscire a fornire sempre una spiegazione sufficientemente esaustiva. Nella maggior parte dei casi, questo tipo di prodotti per il web richiedono infatti all’utilizzatore uno sforzo ulteriore prima di arrivare ad ottenere il suggerimento desiderato. Sarebbe utile, ad esempio, che all’utilizzatore del diagramma sia fornita un’informazione più completa su cos’è un grafico a torta piuttosto che limitarsi a visualizzarne una riproduzione sintetica.

Il Wizard Grafico è stato studiato per rispondere a una seconda e a una terza domanda centrale. La **seconda domanda centrale** ha in realtà la funzione di arricchire/completare l’informazione fornita dal diagramma di

Agarwal. Supponiamo ad esempio che all'utilizzatore sia stato suggerito d'impiegare un grafico a torta. A questo punto la seconda domanda centrale sarà:

Cos'è un grafico a torta?

Il Wizard Grafico fornisce una risposta esaustiva e allo stesso tempo sintetica alla seconda domanda centrale. Per ognuna delle combinazioni possibili di **tipo di rappresentazione/natura dei dati** il Wizard Grafico mette infatti a disposizione, oltre al nome e a un'immagine del grafico, anche una breve spiegazione testuale che aiuta a capire ciò che la sua semplice riproduzione esemplare da sola non può fare.

Tuttavia il principale scopo del Wizard Grafico è rispondere alla **terza domanda centrale**, che sulla base dell'esempio a cui abbiamo accennato in precedenza, sarà:

Come posso costruire un grafico a torta?

La risposta a questo quesito rappresenta per me il vero motivo per cui è stato realizzato il Wizard Grafico. Se da un lato è stato necessario individuare un metodo per guidare l'utilizzatore attraverso il percorso logico per la scelta del grafico più adatto ai propri scopi, dall'altro lato il Wizard Grafico si propone di suggerire alcuni degli strumenti utilizzabili per realizzare il grafico desiderato.

Software basati sul web

Alla base delle scelta degli strumenti suggeriti, vi è la precisa intenzione di limitare le opzioni ai soli **software basati sul web** (*web-based*). La proliferazione di piattaforme web che offrono la possibilità di creare rappresentazioni grafiche e statistiche di alta qualità è tale da consentire ormai a chiunque di ottenere visualizzazioni di base e complesse con pochi clic e senza spendere un euro in costose licenze di utilizzo. Inoltre tali strumenti web, per funzionare, appoggiano spesso su interfacce per l'inserimento dati (*data entry*) altrettanto agevoli da utilizzare e anch'esse basate sul web.

Naturalmente il mondo del web, proprio per la sua caratteristica varietà, comporta una diversità di funzionamento e utilizzo a seconda dello strumento utilizzato. Potrebbe infatti accadere che più software basati sul web impiegabili per creare la medesima rappresentazione richiedano modalità di utilizzo molto differenti. Oltre a ciò, gli strumenti differiscono tra loro in termini di numero di opzioni dedicate all'interattività, di modalità di personalizzazione dell'aspetto e dei contenuti, di qualità stessa della "resa grafica" delle rappresentazioni, eccetera.

Tuttavia, ciò che più conta è il contesto in cui tali strumenti nascono e si sviluppano: il loro utilizzo viene spesso indicato come ideale in contesti di utilizzo di dati aperti (**open data**), in occasioni, cioè, nelle quali sia possibile sfruttare informazioni disponibili per chiunque e in formati universalmente utilizzabili. In alcuni casi questi strumenti sono il frutto di un lavoro collaborativo svolto da utenti-sviluppatori che si muovono in contesti *open source* (r-project), oppure sono il risultato di ricerche accademiche (D3.js), in altri casi ancora si tratta di strumenti implementati in veri e propri portali dedicati alla condivisione dei dati e dei grafici (Many Eyes).

In tutti questi casi, si tratta di soluzioni ideate principalmente per:

1. agevolare una maggiore diffusione delle informazioni;
2. favorire la creazione di grafici di bassa, media e alta complessità;
3. garantire una migliore comprensione del dato numerico, grazie ad una significativa semplificazione dei dettagli grafici.

È dal punto 3, infine, che desidero partire per una considerazione finale: la semplificazione dei grafici richiede un passaggio spesso doloroso ma indispensabile. La sempre maggiore disponibilità di dati numerici in tutti i contesti (marketing, ricerca, divulgazione, ecc.) comporta necessariamente l'impiego di grafici che riescano a spiegare al meglio i risultati delle proprie elaborazioni. Gli strumenti devono quindi essere facili da usare, essere intuitivi, e soprattutto fornire un'ampia gamma di soluzioni di rappresentazione per il medesimo risultato numerico-quantitativo. Allo stesso tempo all'utilizzatore è richiesto di compiere lo sforzo di comprendere cosa desidera ottenere da tali strumenti sulla base dei propri dati e soprattutto come. Tale situazione di pratica continua (nel senso di non saltuaria) favorisce quindi, non solo l'evoluzione degli strumenti di visualizzazione *web-based*, ma anche la nascita di **nuovi metodi di rappresentazione**, attraverso il contributo conoscitivo di tutti gli utilizzatori. Si pensi semplicemente alla difficoltà che chiunque avrebbe incontrato nel cercare di descrivere graficamente un fenomeno a più di due dimensioni (variabili) solo qualche decennio fa, in modo efficace e sufficientemente comprensibile. Oggi creare con pochi clic un complesso grafico tridimensionale è un'operazione praticamente alla portata di tutti. Oggi chiunque può creare raffigurazioni animate dell'evoluzione dei propri dati storici solamente conoscendo in quale forma utilizzare i dati attraverso interfacce simili a veri e propri fogli elettronici (Gapminder). Oggi creare grafici inerenti a contesti molto specialistici per consegnarli alla comunità degli utilizzatori come

strumenti di riferimento di rappresentazione è un'operazione relativamente semplice rispetto a non molto tempo fa. La maggiore conoscenza degli strumenti utilizzabili, la loro maggiore diffusione ed il loro contemporaneo utilizzo favorisce quindi anche un processo creativo indispensabile a garantire sempre maggiori possibilità di sviluppo e di varietà di scelta.

Il Wizard Grafico ha anche questo scopo: ovvero di stimolare l'inventiva nel suo utilizzatore una volta trasmessagli la concreta sensazione di poter realizzare ex novo nuove forme di rappresentazione. È mio auspicio, infatti, che i suggerimenti offerti dal Wizard Grafico non si limitino a fornire ai loro utilizzatori quanto necessario per ottenere la raffigurazione ricercata, ma soprattutto finiscano per rappresentare un elemento di fascino tale da indurre ad immaginare in quali e quanti altri modi non ancora esistenti descrivere i propri dati.

Duccio Schiavon

Capitolo 2

Relazione

Scatterplot

Binning

Curva di adattamento

Heatmap

Bubble chart

Superficie

Linee di livello

Matrice di correlazione

Matrici di grafici

2

Relazione tra due variabili quantitative

Lo **scatterplot** è uno strumento grafico attraverso il quale associare due variabili quantitative (continue o discrete). Viene principalmente utilizzato per dedurre se vi sono relazioni di tipo direttamente o inversamente proporzionale tra le due misure confrontate. Si tratta inoltre di un grafico utile qualora si desideri dedurre attraverso un'unica visualizzazione la natura distributiva delle due misure.

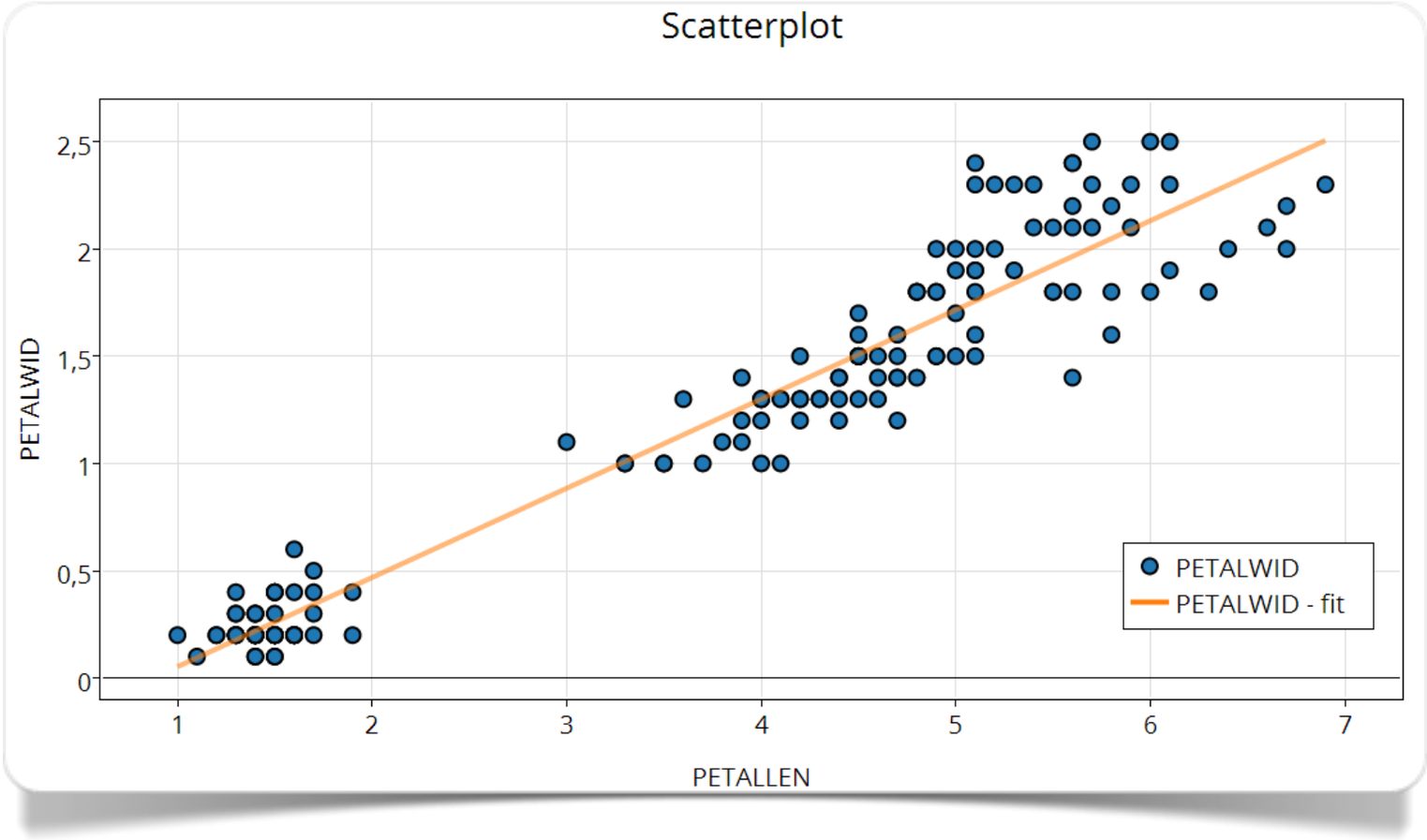


Fig. 2.1 Scatterplot creato con plotly

plotly è un portale web per la produzione di grafici statistici interattivi, assolutamente responsive (si adattano alle dimensioni dello schermo, e per questo garantiscono la compatibilità mobile), provvisto di API per l'interfacciamento con ambienti esterni e di un'ampia gamma di possibilità grafiche. Attraverso la sua interfaccia è possibile caricare file in formato CSV, TSV, Matlab, MS Access, o testo, per poi esportare i suoi grafici in formato PNG, PDF, SVG ed EPS. Inoltre è possibile interfacciarsi ad R attraverso la sua API.

Il **binning** è un tipo di rappresentazione che consente d'individuare visivamente le aree più "popolate" di uno **scatterplot**. In termini tecnici, il *data binning* è una tecnica di trattamento dei dati in cui i valori originali che cadono in un dato intervallo minimo (*bin*) sono sostituiti da un valore unico rappresentativo di tale intervallo, spesso rappresentato dal valore "centrale". Come riferimento riguardante questa particolare rappresentazione grafica, si veda il lavoro di Zachary Forest Johnson.

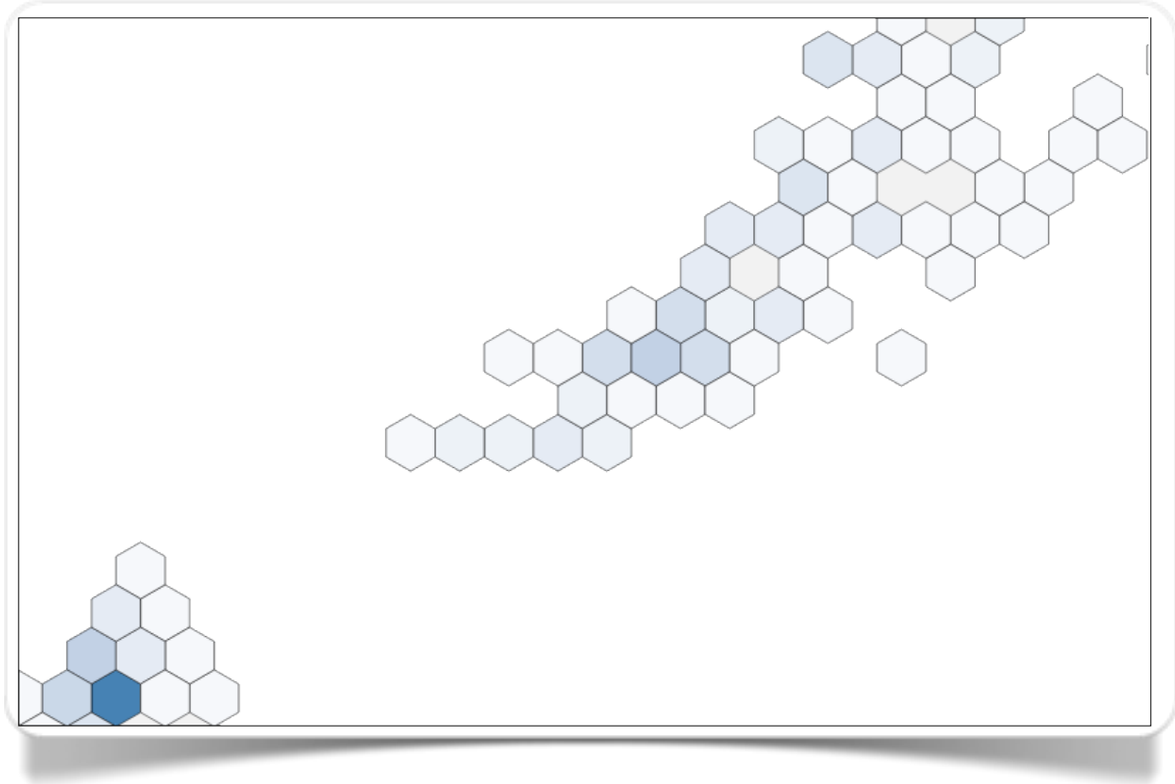


Fig. 2.2 Hexagonal binned scatterplot creato con Raw

Raw è una piattaforma *open source*, la cui interfaccia web funziona da generatore di visualizzazioni. Il suo funzionamento prevede il caricamento dei propri dati attraverso un semplice copia-incolla all'interno di un'area di testo, la selezione del layout da utilizzare, la specificazione delle variabili da utilizzare, la personalizzazione di alcuni aspetti grafici, ed

infine l'esportazione del grafico in semplice formato PNG oppure anche in formato vettoriale SVG. Si basa sulla libreria D3.js.

La **curva di adattamento** viene impiegata in particolare in presenza di variabili quantitative con molti dati di **natura continua**. È molto utile per stabilire le tendenze delineate dalla relazione di due variabili a confronto, e per valutare il livello di scostamento dei punti dato dalla **curva interpolante** (**variabilità**).

WolframAlpha è un motore computazionale in grado di elaborare le parole chiave specificate dall'utente e di fornire una serie d'informazioni numeriche, dati e informazioni. Gli sviluppatori di questo motore di ricerca sono i medesimi che hanno sviluppato il software *Mathematica*: questo il motivo del suo forte orientamento al calcolo e alla statistica. La curva di adattamento in figura 2.3 è stata realizzata specificando nel suo campo di ricerca l'espressione:

```
exponential fit
0.783,0.552,0.383,0.245,0.165,0.097
```

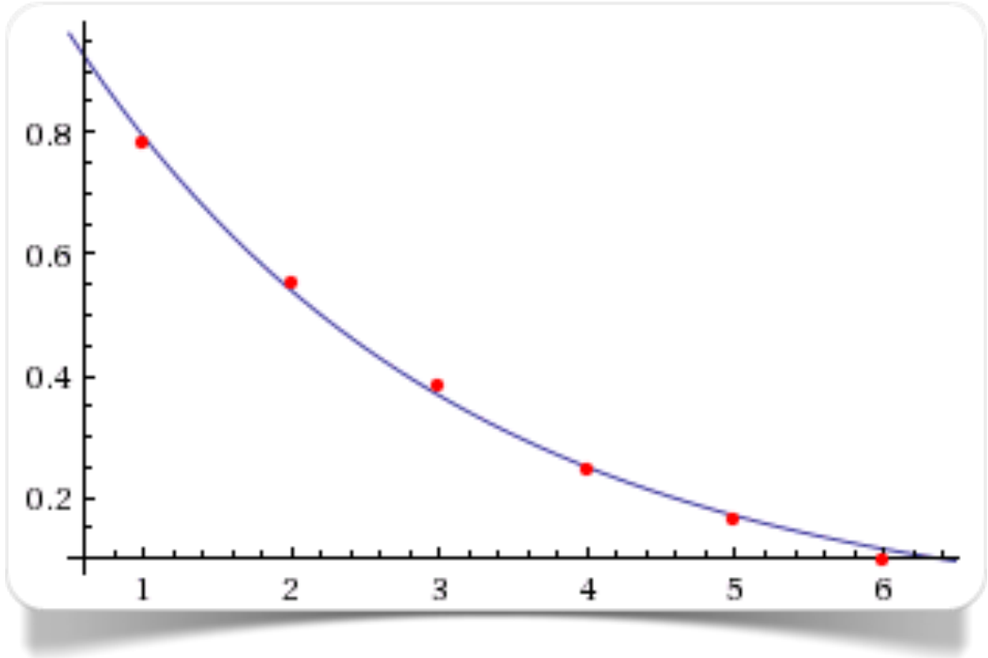


Fig. 2.3 Curva di adattamento realizzata con WolframAlpha

Tra le possibili alternative, per la creazione si curve di adattamento si consiglia l'utilizzo di plotly, il quale, a partire da uno scatterplot, offre la possibilità di utilizzare l'opzione FIT DATA per adattare una qualsiasi funzione ai punti riportati sul grafico.

In figura 2.4 possiamo osservare uno **scatterplot** in cui è rappresenta la relazione tra speranza di vita e durata delle gestazione in alcune specie di animali.

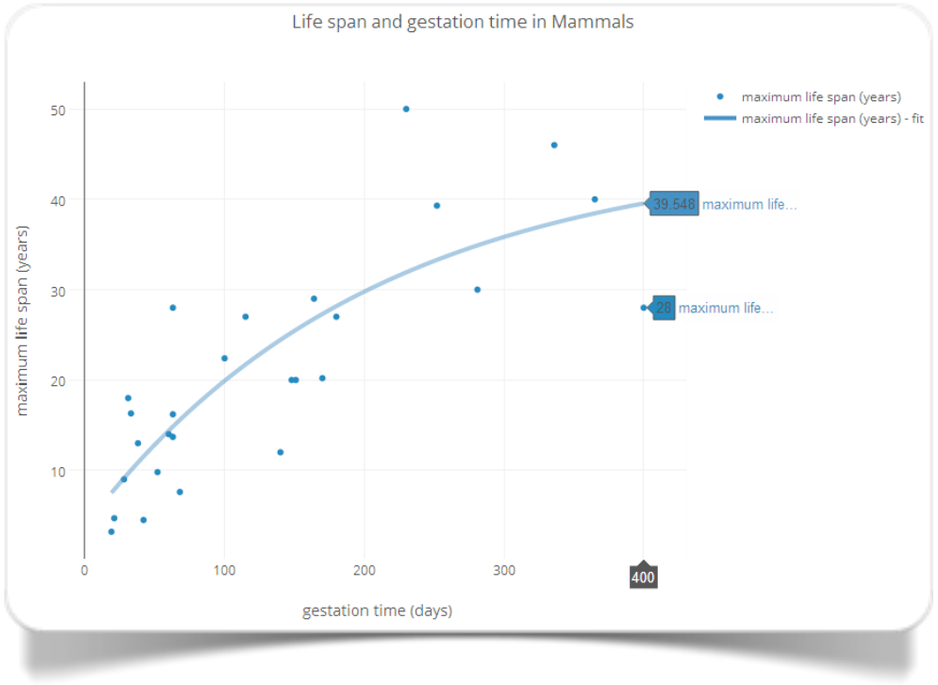


Fig. 2.4 Speranza di vita e periodo di gestazione in alcune specie di mammiferi (elab. plotly)

Species of Animals	maximum life span (years)	gestation time (days)
African-giant-pouched-rat	4.5	42
Arctic-Fox	14	60
Baboon	27	180
Cat	28	63
Chimpanzee	50	230
Cow	30	281
Donkey	40	365
Giraffe	28	400
Goat	20	148
Gorilla	39.3	252
Gray-wolf	16.2	63
Ground-squirrel	9	28
Guinea-pig	7.6	68
Horse	46	336

Fig. 2.5 Dati relativi alla figura 2.4

Relazione tra due variabili qualitative

Il grafico di tipo **heatmap** (Sneath, 1957) è la riproduzione visiva ideale di una **tabella di contingenza** a doppia entrata: attraverso di esso vengono confrontate due variabili categoriali caratterizzate da un numero limitato di categorie. La gradazione dei colori è indicativa della dimensione delle frequenze di ogni cella. A valori (ad es., percentuali) più grandi corrisponderanno colori più intensi.

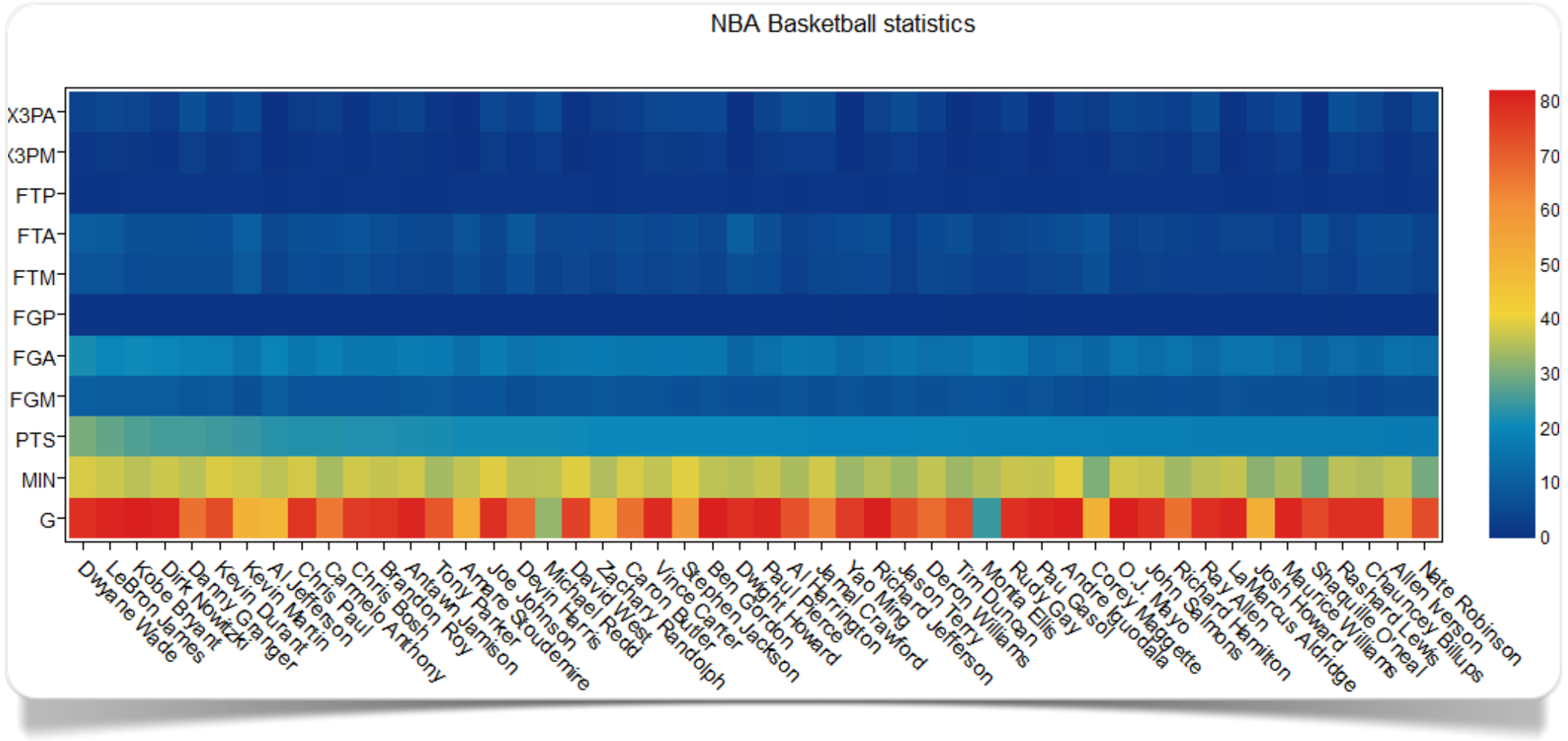


Fig. 2.6 Heatmap realizzata con plotly

Grazie a plotly è possibile creare **heatmap**, con possibilità di controllo su tutti i suoi aspetti grafici. Nella *gallery* di visualizzazioni riproducibili è possibile trovarne degli esempi come quello della figura 2.6.

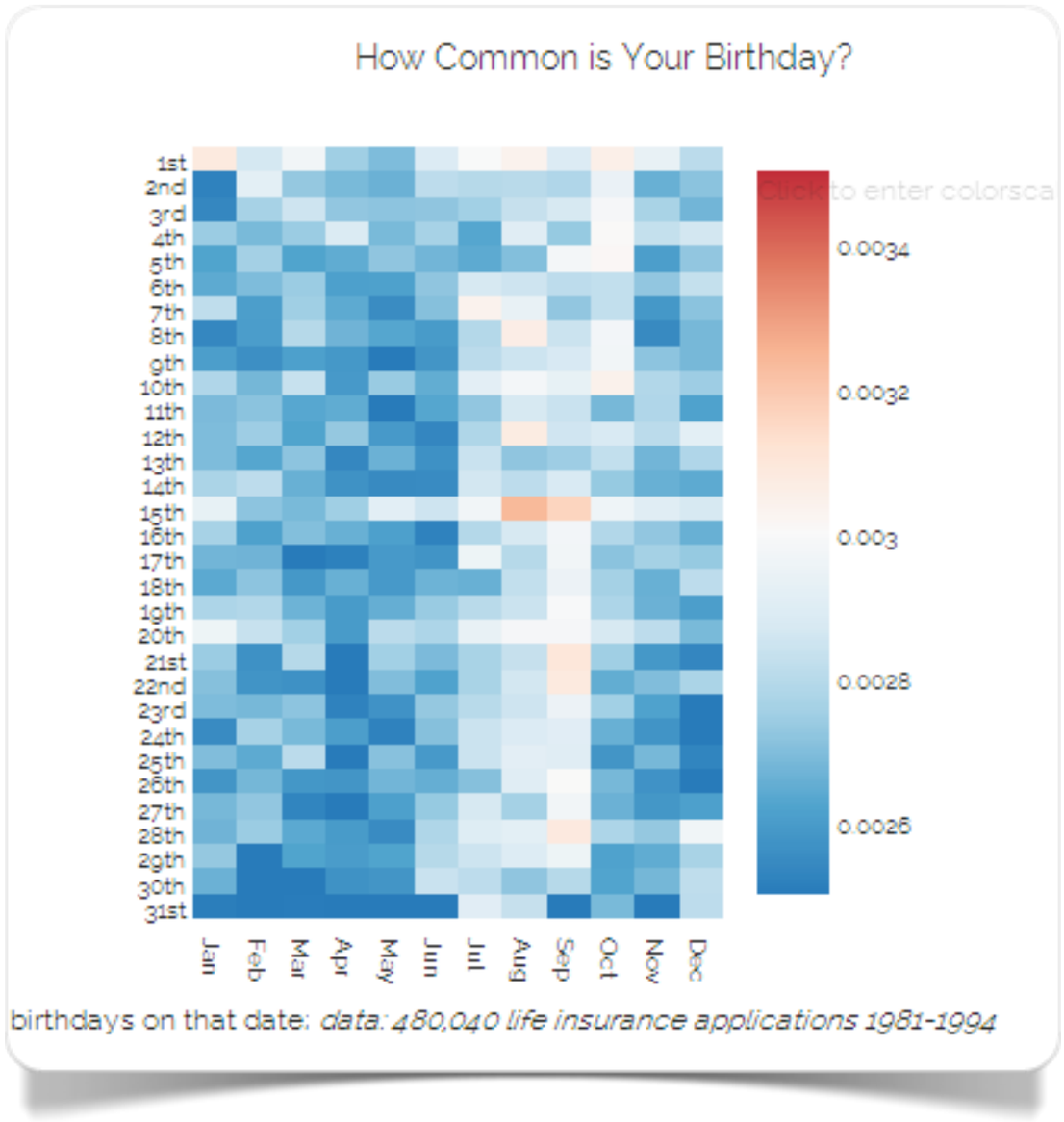


Fig. 2.7 Frequenza della data di nascita per giorno e mese (elab. di Dreamshot con plotly

In figura 2.7 possiamo osservare come la data di nascita più comune, nella popolazione scelta dall'utente Dreamshot che ha realizzato il grafico con plotly, sia il 15 agosto. Il grafico mette in evidenza come nei mesi estivi le nascite siano più frequenti.

In alternativa, per la creazione di grafici **heatmap**, può essere utilizzato Many Eyes, come in questo esempio.

Relazione tra tre variabili quantitative

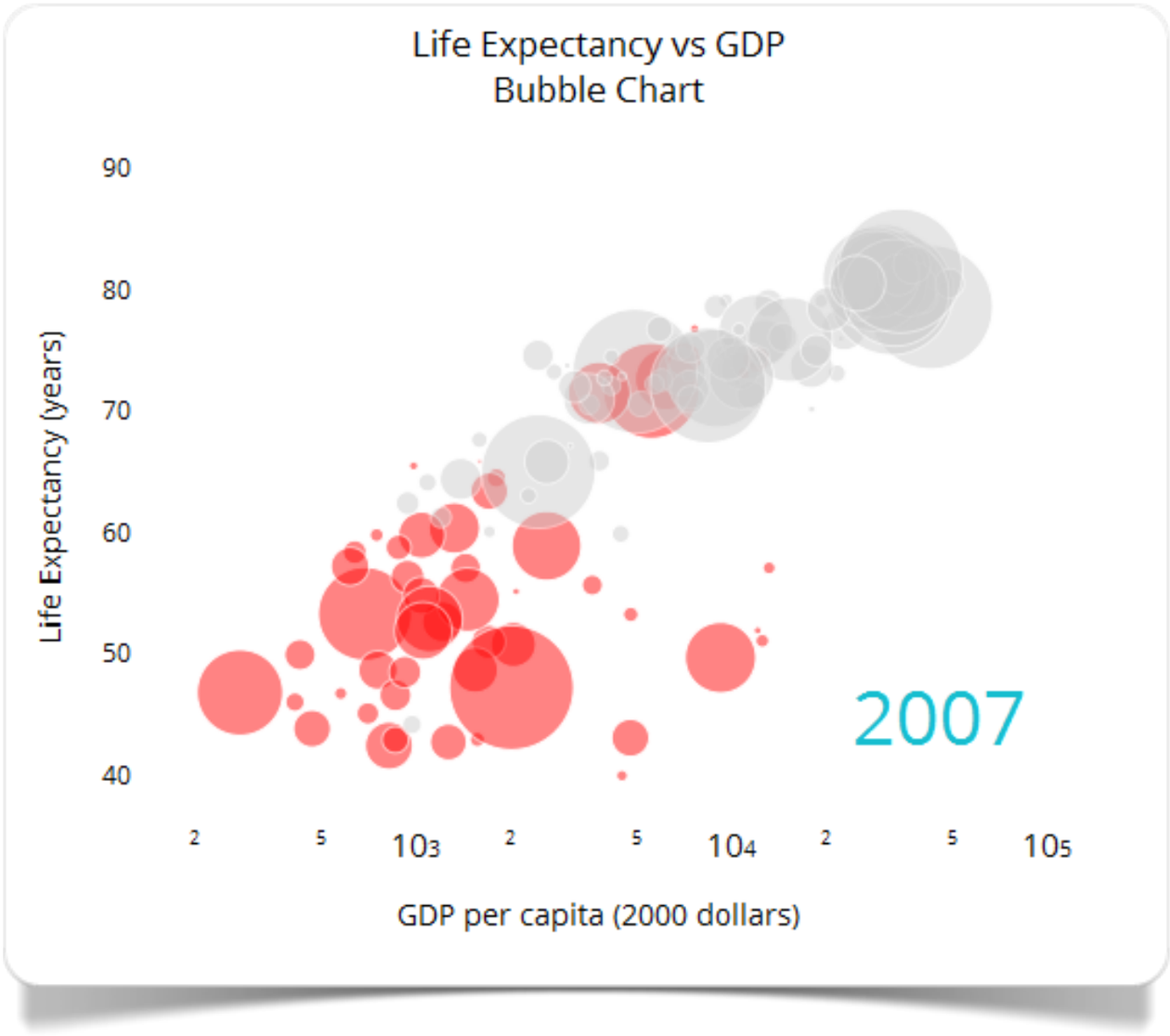


Fig. 2.8 Bubble chart realizzato con plotly

Lo **scatterplot** (Chambers 1983), oltre a consentire di associare due variabili quantitative per determinare se vi sia un rapporto di proporzionalità diretta o inversa tra di esse, permette di considerare opzionalmente una terza variabile “di entità” (Z). Questa variabile, anch’essa numerica, definisce l’ordine di grandezza di ogni singolo punto-dato all’interno del grafico. Per distinguerlo da un semplice scatterplot a due dimensioni questo tipo di rappresentazione è spesso chiamato anche **bubble chart**.

In figura 2.8 è riportato un bubble chart realizzato con plotly. E’ possibile realizzare questo tipo di grafico anche usando Raw oppure Many Eyes.

Il grafico di figura 2.9 è stato realizzato con Many Eyes (modalità “scatterplot”) mentre quello di figura 2.10 è stato realizzato con Raw. Il set di dati è identico; il grafico 2.10 utilizza il subset ridotto di dati riportato nella finestra a scorrimento della figura 2.11.

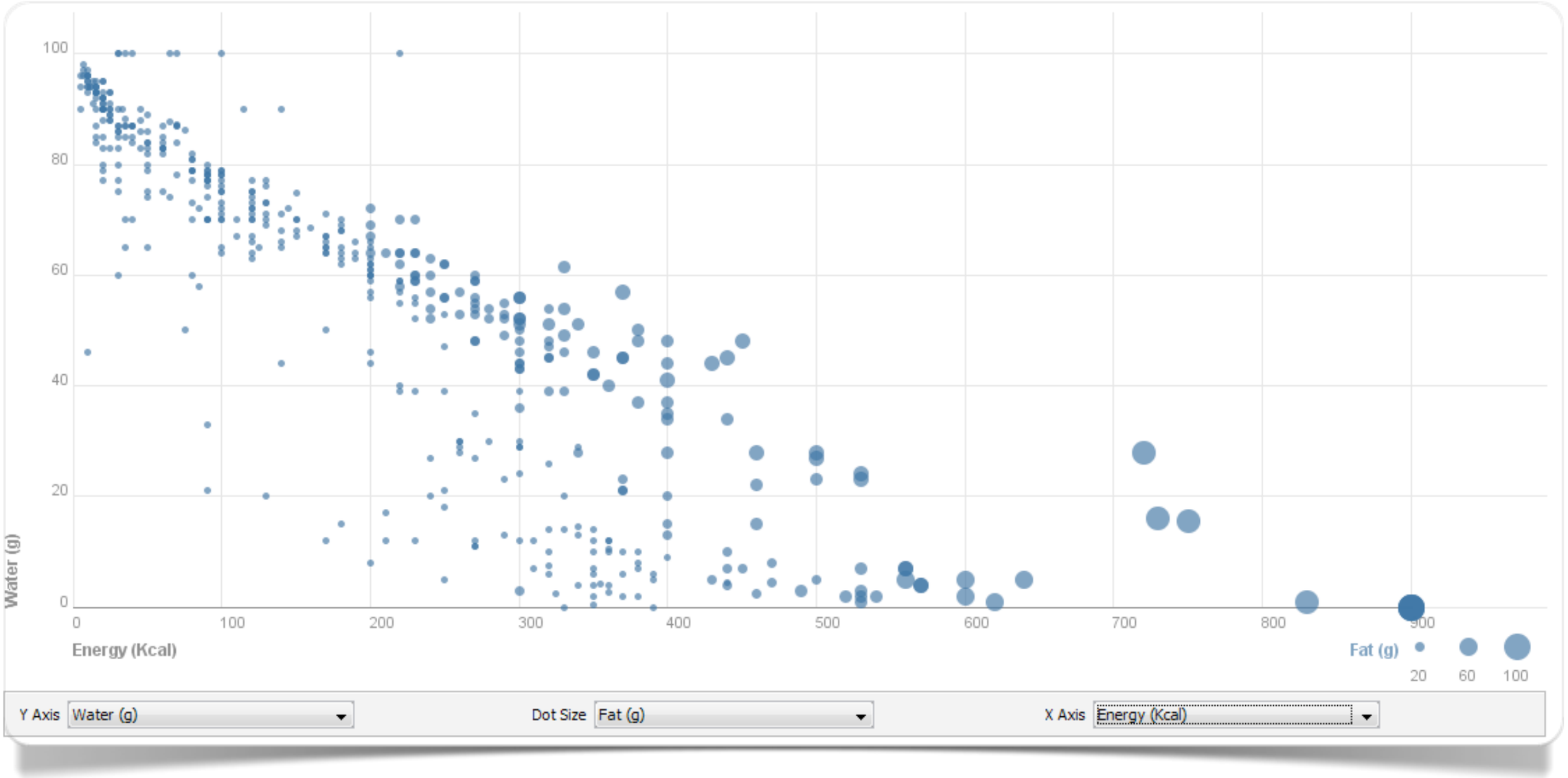


Fig. 2.9 Alcuni tipi di alimenti e loro proprietà nutrizionali (elab. Many Eyes)

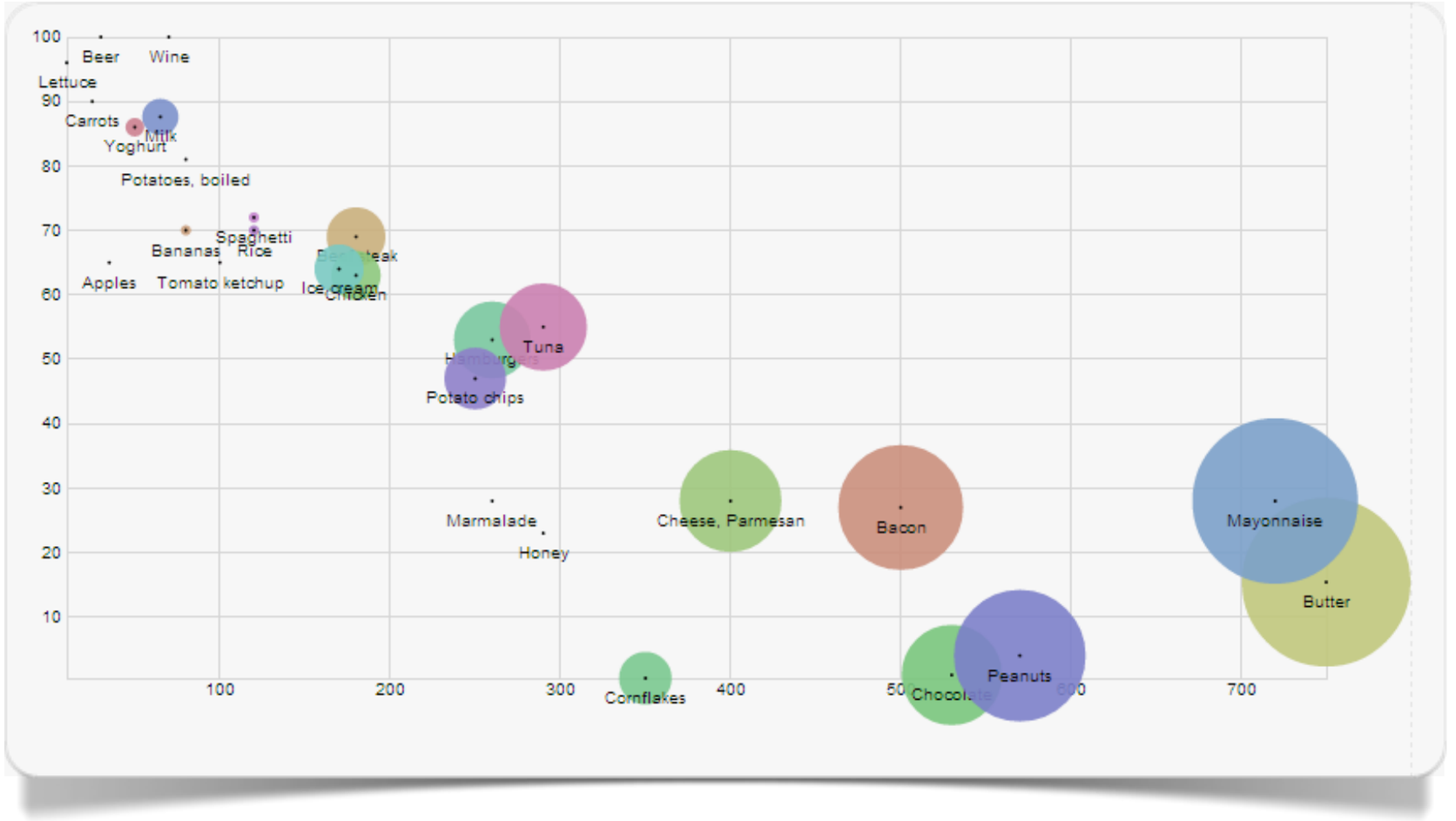


Fig. 2.10 Relazione tra Calorie (x), Acqua (y) e Grassi (z) per 100 g. di alcuni alimenti (elab. Raw)

Type of Food (100g)	Energy (Kcal)	Protein (g)	Fat (g)	Water (g)	Vitamin A (mg)	Vitamin B1 (mg)
Apples	35	0.2	0	65	0	0.03
Bacon	500	23	45	27	0	0.4
Bananas	80	1	0.3	70	200	0.04
Beef steak	180	20	10	69	0	0.06
Beer	30	0	0	100	0	0
Butter	750	0.5	82	15.4	1000	0
Carrots	25	0.7	0	90	12000	0.06

Fig. 2.11 Alcuni tipi di alimenti e loro proprietà nutrizionali

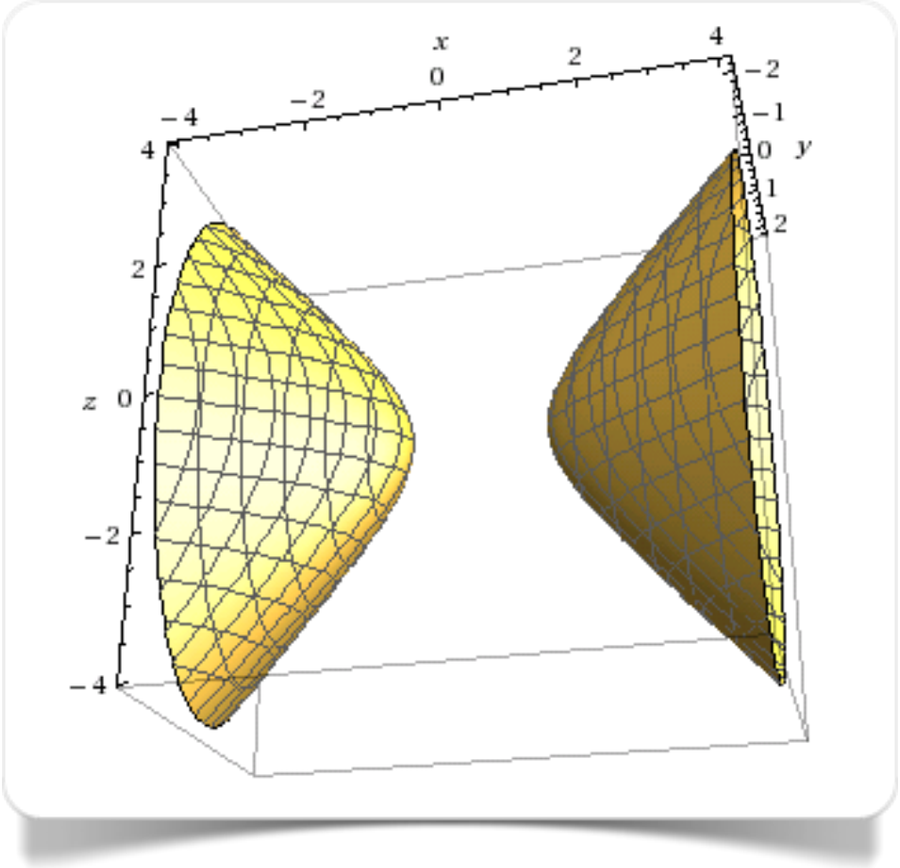


Fig. 2.12 Grafico 3D prodotto con WolframAlpha

La **superficie** è un particolare tipo di grafico che consente la rappresentazione tridimensionale di tre variabili quantitative (continue, soprattutto, ma pure ordinali se necessario).

Questi grafici hanno il particolare pregio di sfruttare diversi elementi visivi. Come in una carta topografica, ad esempio, i colori e i motivi servono per rappresentare le aree che contengono lo stesso intervallo di valori.

WolframAlpha supporta la creazione di grafici 3D attraverso la semplice specificazione di una formula all'interno del campo di ricerca: il grafico in figura 2.11 è stato creato inserendo nel campo di ricerca l'espressione :

```
plot x^2 - 3y^2 - z^2
```

Il **grafico a linee di livello** (meglio noto come *contour plot*) è la perfetta trasposizione bidimensionale di un grafico a superficie. Una volta stabilite le variabili da riportare sulle assi dell'ascissa e dell'ordinata, la terza variabile sarà rappresentata da linee e curve riportate sul piano dimensionale. Ogni intervallo definito dallo spazio incluso tra le diverse curve rappresenta una particolare classe di variazione dei valori della variabile Z, contrassegnata a sua volta da un particolare colore di gradiente.

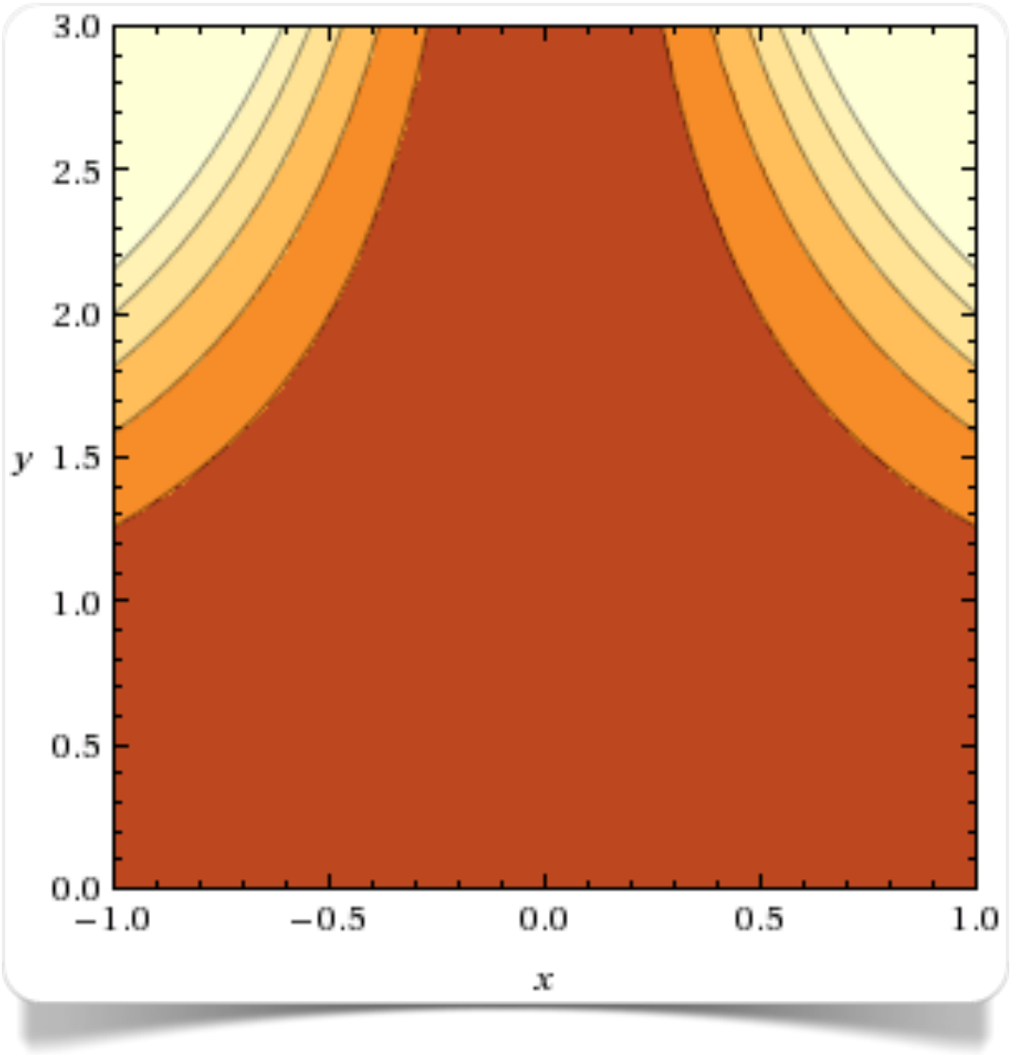


Fig. 2.13 Grafico a linee di livello realizzato con WolframAlpha

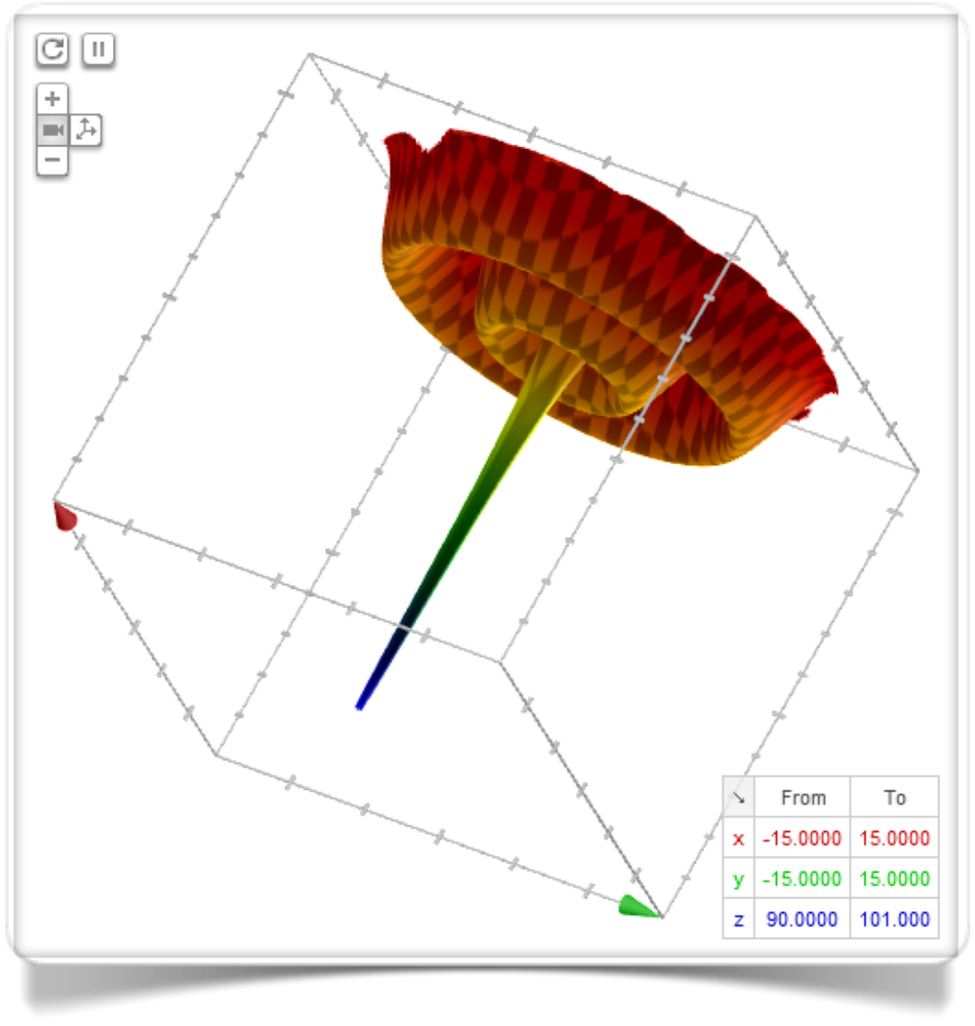


Fig. 2.14 Grafico 3D realizzato con Google Search

Il grafico a linee di livello di figura 2.12 è stato realizzato con WolframAlpha specificando nel suo campo di ricerca l'espressione

```
plot x^2 y^3, x=-1..1, y=0..3
```

Solitamente i grafici a linee di livello vengono prodotti dall'applicativo a fianco della rispettiva rappresentazione 3D con grafico a superficie.

Anche Google supporta la creazione di grafici 3D attraverso la semplice specificazione di una formula all'interno del campo di ricerca: il grafico di figura 2.14 è stato creato inserendo nel campo di ricerca l'espressione :

```
100-3/(sqrt(x^2+y^2))+sin(sqrt(x^2+y^2))+sqrt(200-(x^2+y^2)+10*sin(x)+10sin(y))/1000,
                x is from -15 to 15, y is from -15 to 15, z is from 90 to 101
```

La particolare tecnologia WebGL, su cui si basa la funzione di resa grafica 3D di Google consente di utilizzare alcune interessanti opzioni d'interattività quale ad esempio la funzione di zoom o il trascinamento del grafico con rotazione lungo uno dei tre assi.

Relazione tra molte variabili quantitative

La **matrice di correlazione** è rappresentata da una matrice quadrata NxN, organizzata in modo che le righe corrispondano ad N variabili quantitative. Le colonne a loro volta devono corrispondere alle medesime N variabili quantitative, ordinate secondo il medesimo ordine consequenziale in cui queste sono organizzate per riga. La matrice sarà quindi composta da NxN quadrati ognuno colorato su scala cromatica differente a seconda del valore dell'indice di correlazione calcolato per ogni confronto possibile tra coppie di variabili.

plotly consente di costruire una matrice di correlazione sulla base di un qualsiasi dataset composto da N variabili (colonne). plotly calcola automaticamente gli indici di correlazione associandovi un colore su scala cromatica graduata.

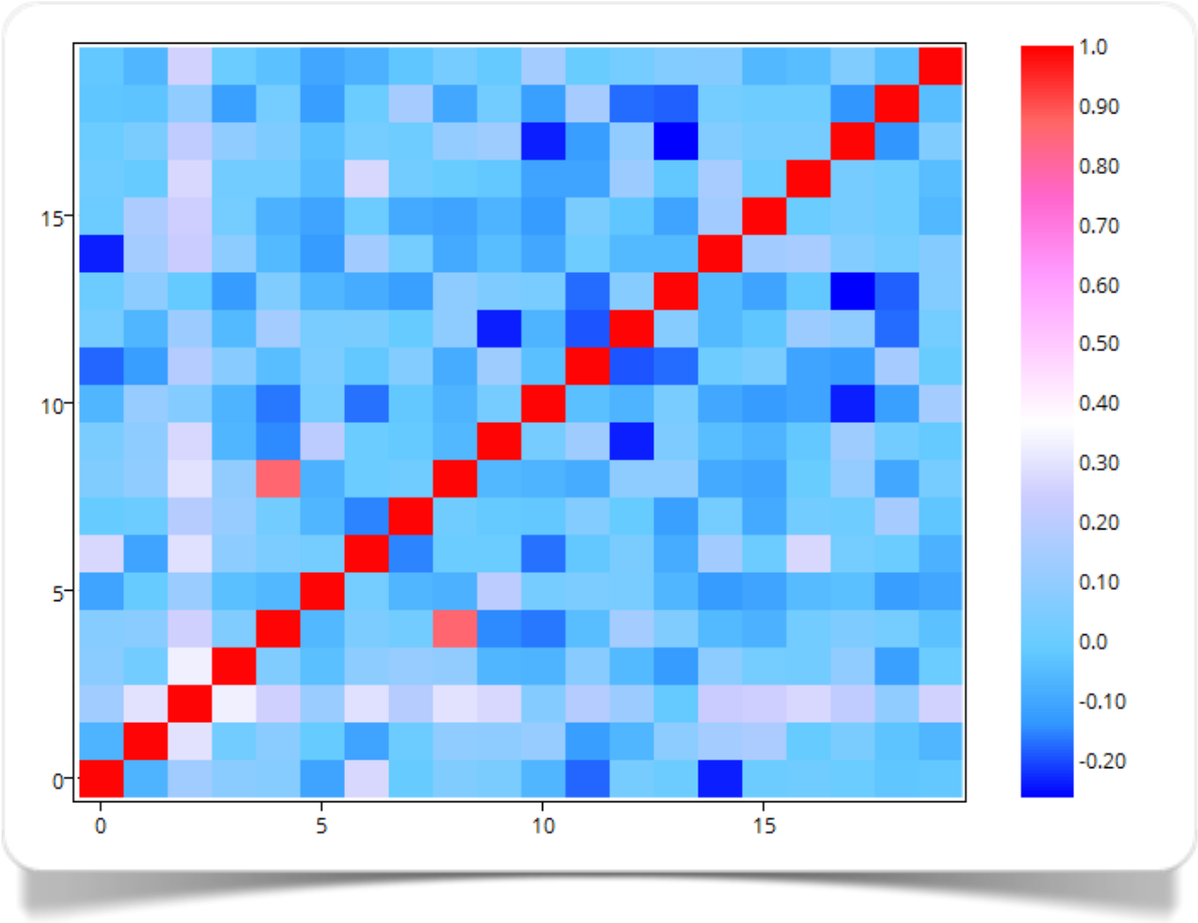


Fig. 2.15 Matrice di correlazione realizzata con plotly

Relazione tra molte variabili

Le **matrici di grafici** consistono in vere e proprie griglie a doppia entrata in cui vengono riportati in corrispondenza di ogni “incrocio” il singolo grafico (scatterplot, grafico a torta, istogramma) relativo al confronto tra coppie di variabili.

Il tipo di visualizzazione Matrix Chart di Many Eyes permette di rappresentare dati multidimensionali all’interno di una vera e propria griglia. In particolare, lo strumento grafico utilizzato per rappresentare ogni cella della griglia può essere un grafico a bolle e a torta.

Fig. 2.16 Matrice di grafici a bolle creata con Many Eyes

Nell'esempio della figura 2.17 vediamo una matrice con grafici a bolle. I dati sono ripresi da una rapporto di Donna J. Nelson e Christopher N. Brammer sulla presenza della donne e delle minoranze nella ricerca universitaria e nelle discipline tecnico-scientifiche. I tre gruppi presi in considerazione nel grafico sono i laureati (BS: bachelors of science), i dottorati di ricerca (PhD) e i professori di ogni grado nelle 50 università più importanti degli Stati Uniti. Il diametro delle bolle è riferito alla percentuale di donne per ciascun settore disciplinare. Si può osservare come vi sia una forte disparità nella presenza di donne tra i laureati e i dottorati in tutte le discipline rispetto alla carriera accademica. Vi sono evidenti differenze tra le discipline scientifiche e le discipline più umanistiche come la sociologia e la psicologia. Nel grafico interattivo è possibile selezionare una rappresentazione con matrici di grafici a barre anziché con bolle.

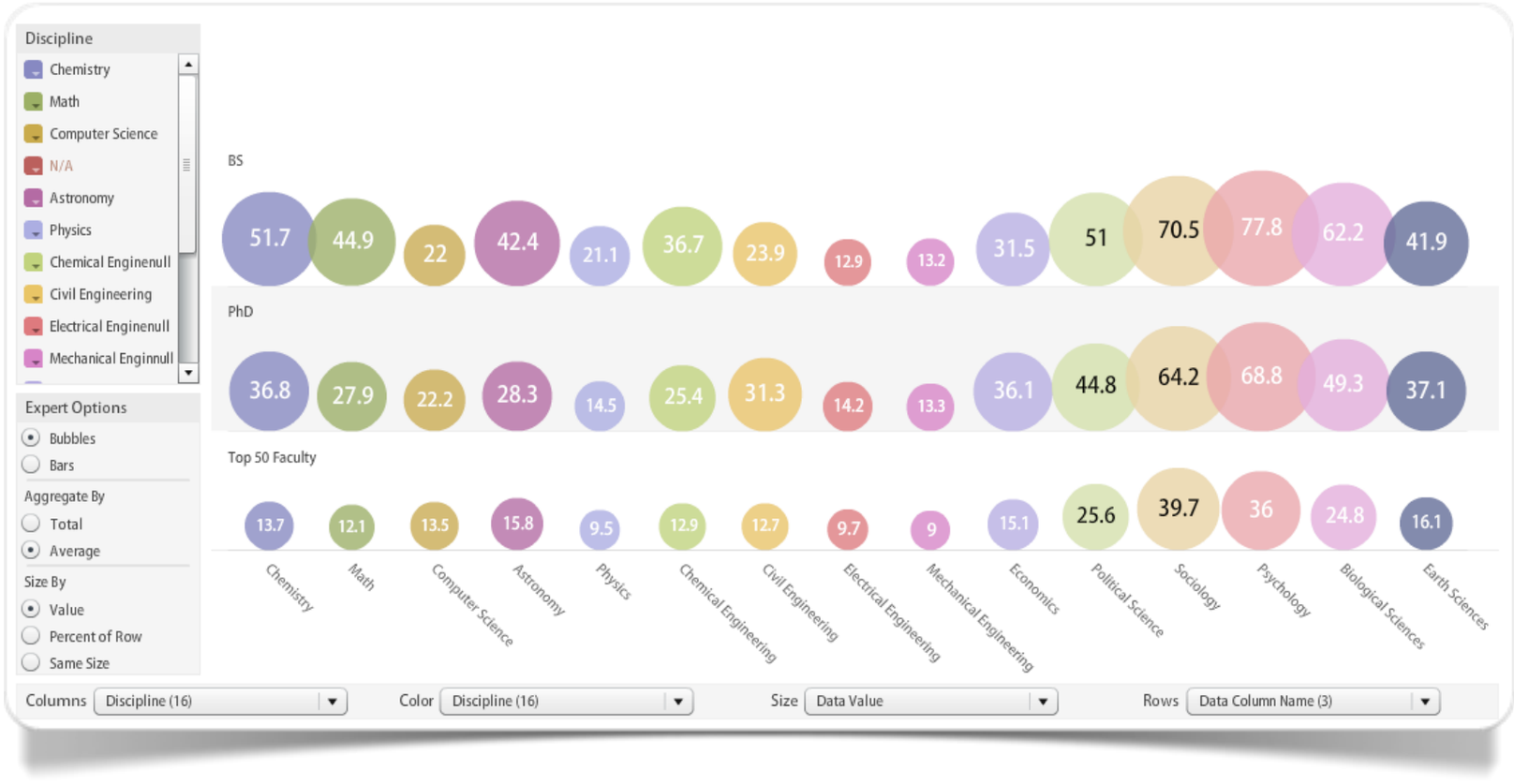


Fig. 2.17 La presenza delle donne in diversi gradi accademici per settore disciplinare (elab. Many Eyes)

Capitolo 3

Distribuzione

Plot dei punti-dato

Area

Istogramma

Istogramma categorizzato

Curva di adattamento

3

Distribuzione di una singola variabile quantitativa

Il semplice **plot dei punti-dato** di una singola **variabile** quantitativa consente di riportare su grafico ogni dato associandovi un marcatore di punto. Solitamente all'interno del plot, nella più semplice delle sue rappresentazioni, i valori dei dati sono riportati sull'asse verticale (Y) mentre sull'asse orizzontale (X) viene riportato il numero di ordinamento corrispondente ai singoli valori.

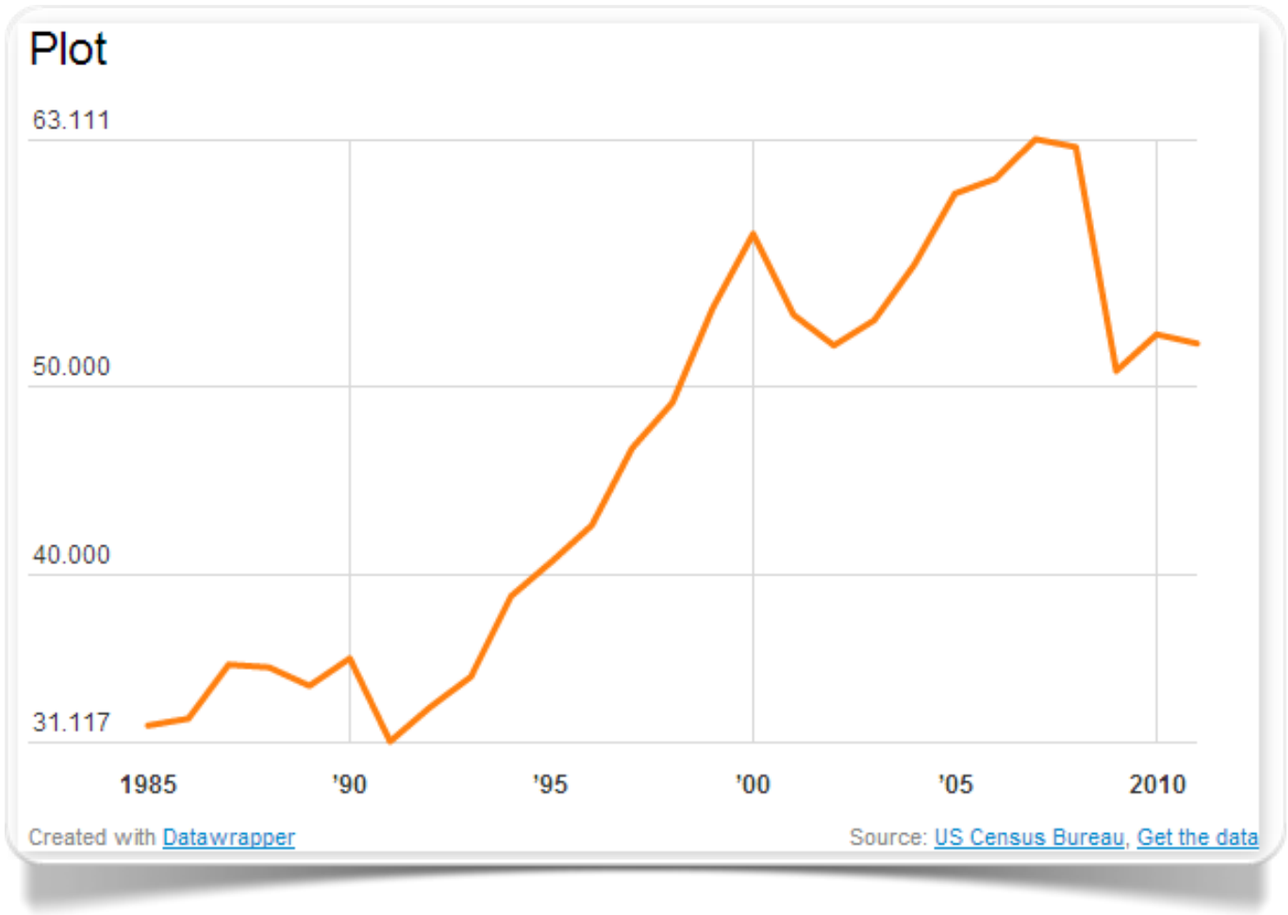


Fig. 3.1 Plot creato con Datawrapper

Tra tutti gli strumenti web per la produzione "al volo" di grafici interattivi e responsive, il più semplice da utilizzare è certamente Datawrapper.
Le fasi di produzione del grafico prevedono 4 step: caricamento dati, verifica e descrizione, visualizzazione, e infine pubblicazione.
Il grafico in figura 3.1 rappresenta l'andamento temporale di una serie univariata di dati.

L'**area** è un tipo di grafico identico al plot ad eccezione dell'area riempita al di sotto della linea con un colore che indica il volume.

In figura 3.2 è possibile osservare un utilizzo ideale dell'area. Trattandosi di rappresentare le quote altimetriche è evidentemente che l'effetto ottenuto tramite il riempimento con il colore, rende maggiormente l'idea figurativa dei picchi delle alture. Il grafico in figura 3.2 è stato realizzato con plotly.

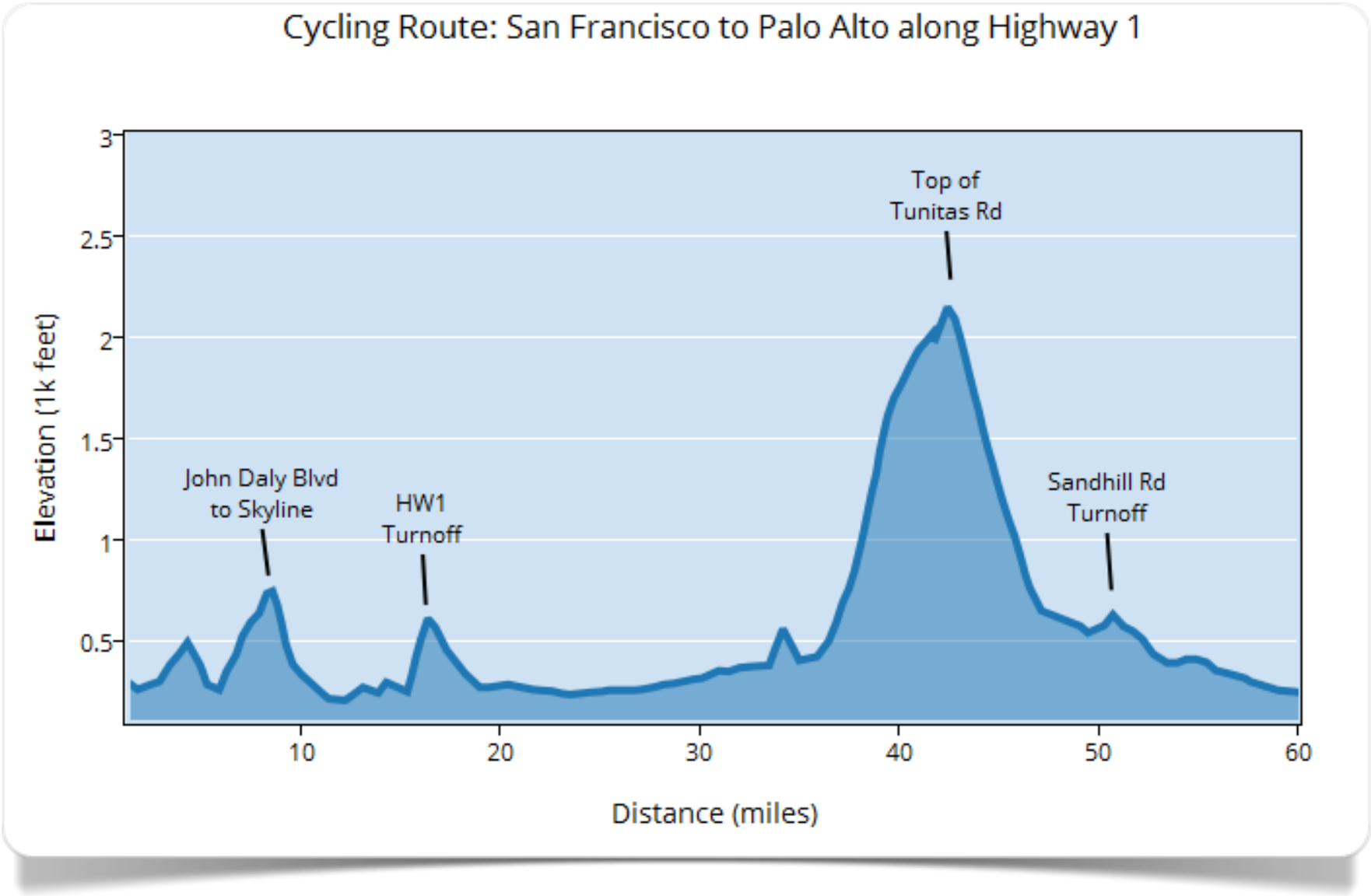


Fig. 3.2 Plot creato con plotly

Distribuzione di una singola variabile (pochi dati)

L'**istogramma** (Pearson, 1895) è un grafico a barre in cui ogni barra rappresenta la frequenza in cui un numero (nel caso di variabili quantitative) o una categoria (nel caso di variabili qualitative) ricorre all'interno della **variabile** considerata. Questo tipo di grafico è particolarmente efficace quando si dispone di un numero limitato di casi. Il grafico in figura 3.3 è stato realizzato con Datawrapper.

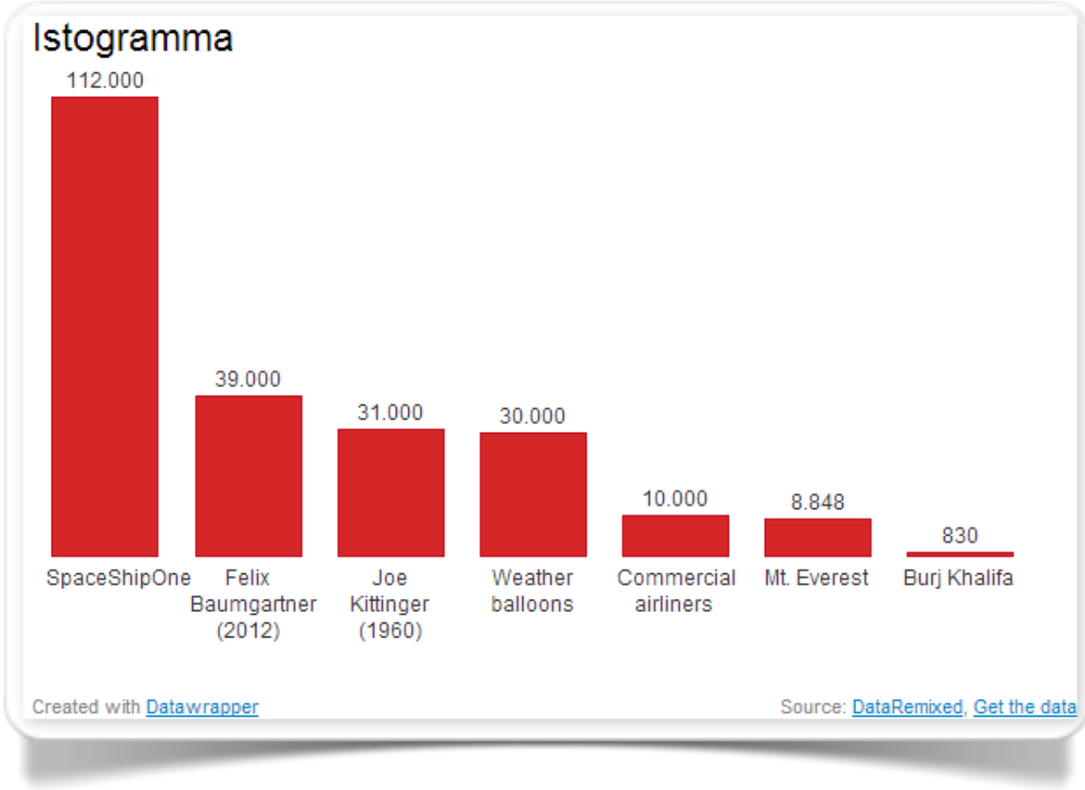


Fig. 3.3 Istogramma creato con Datawrapper

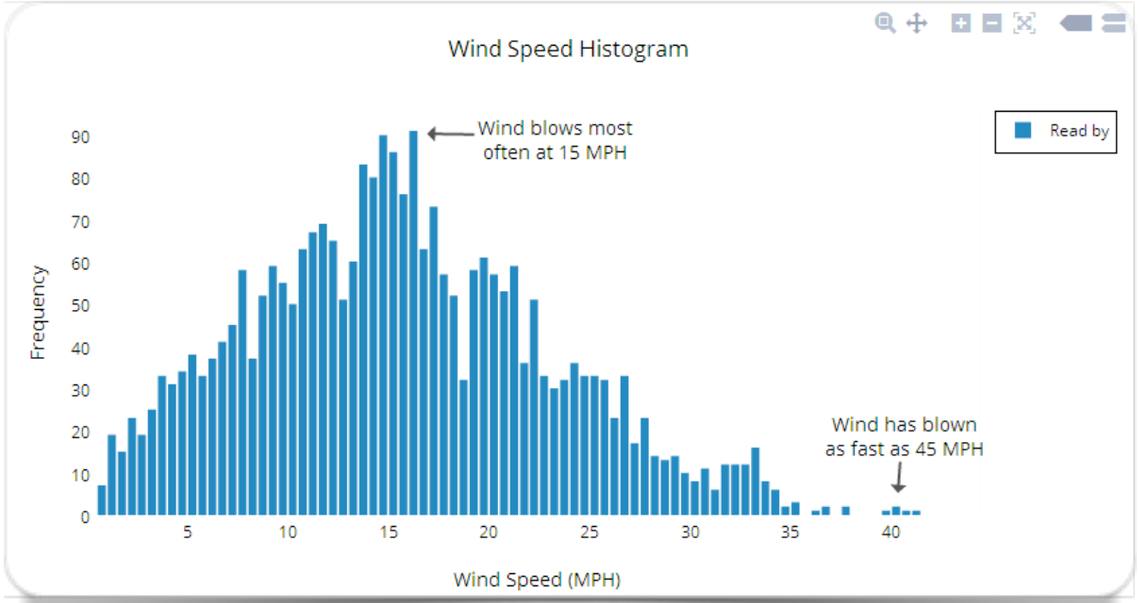


Fig. 3.4 Istogramma realizzato con plotly

Alternative valide per la produzione di istogrammi sono certamente plotly (vedi figura 3.4), che si caratterizza per la possibilità d'inserire ad esempio commenti testuali all'interno all'area del grafico, oppure ancora Many Eyes (vedi figura 3.5) del quale è di particolare interesse esaminare la flessibilità delle interazioni nella versione online.

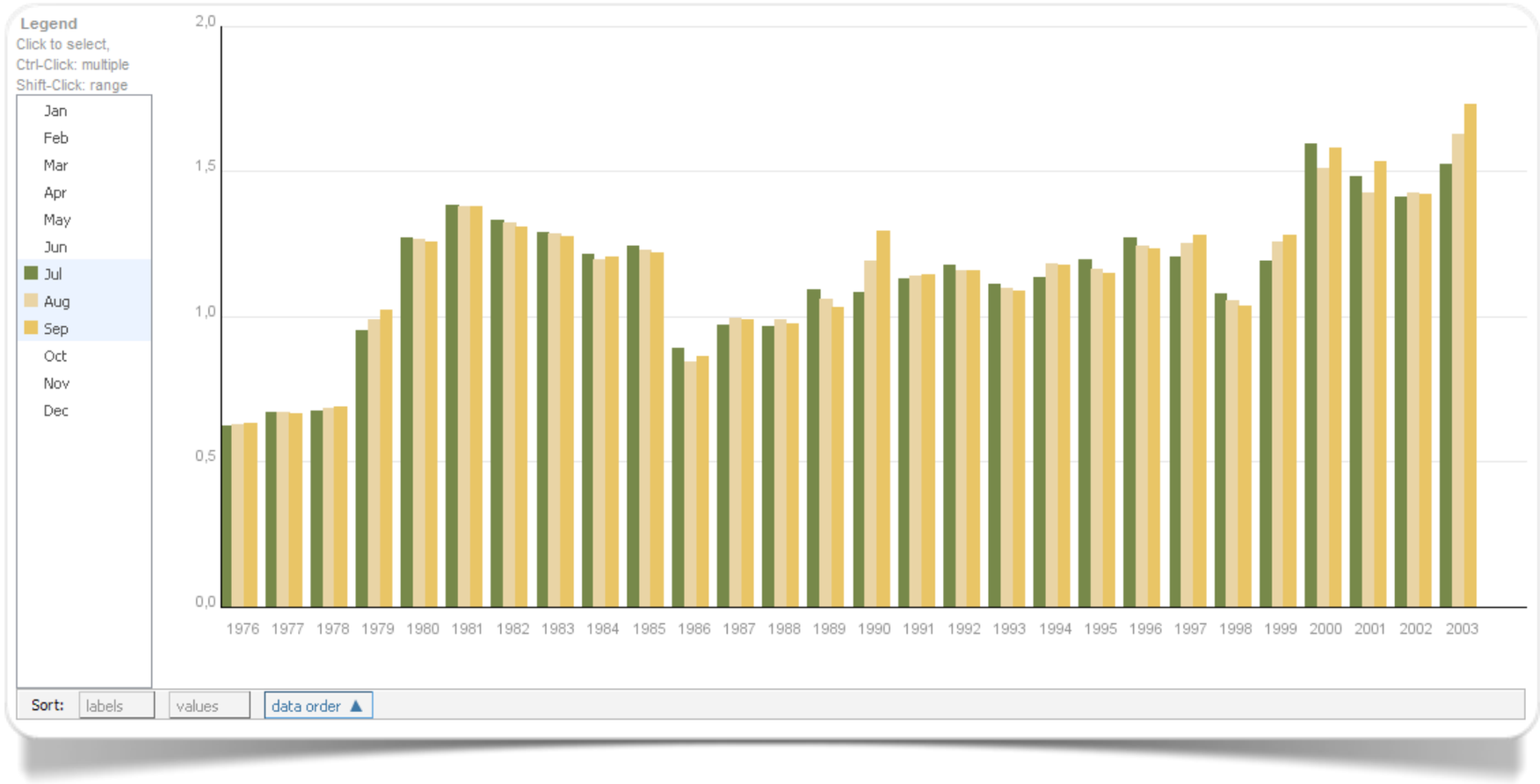


Fig. 3.5 Prezzo del carburante per gallone (elab. Many Eyes)

Distribuzione di più variabili (poche variabili)

L'**istogramma categorizzato** è un istogramma che consente di rappresentare più distribuzioni contemporaneamente. In questi casi si usa utilizzare un colore differente per ognuna delle singole dimensioni coinvolte nel confronto.

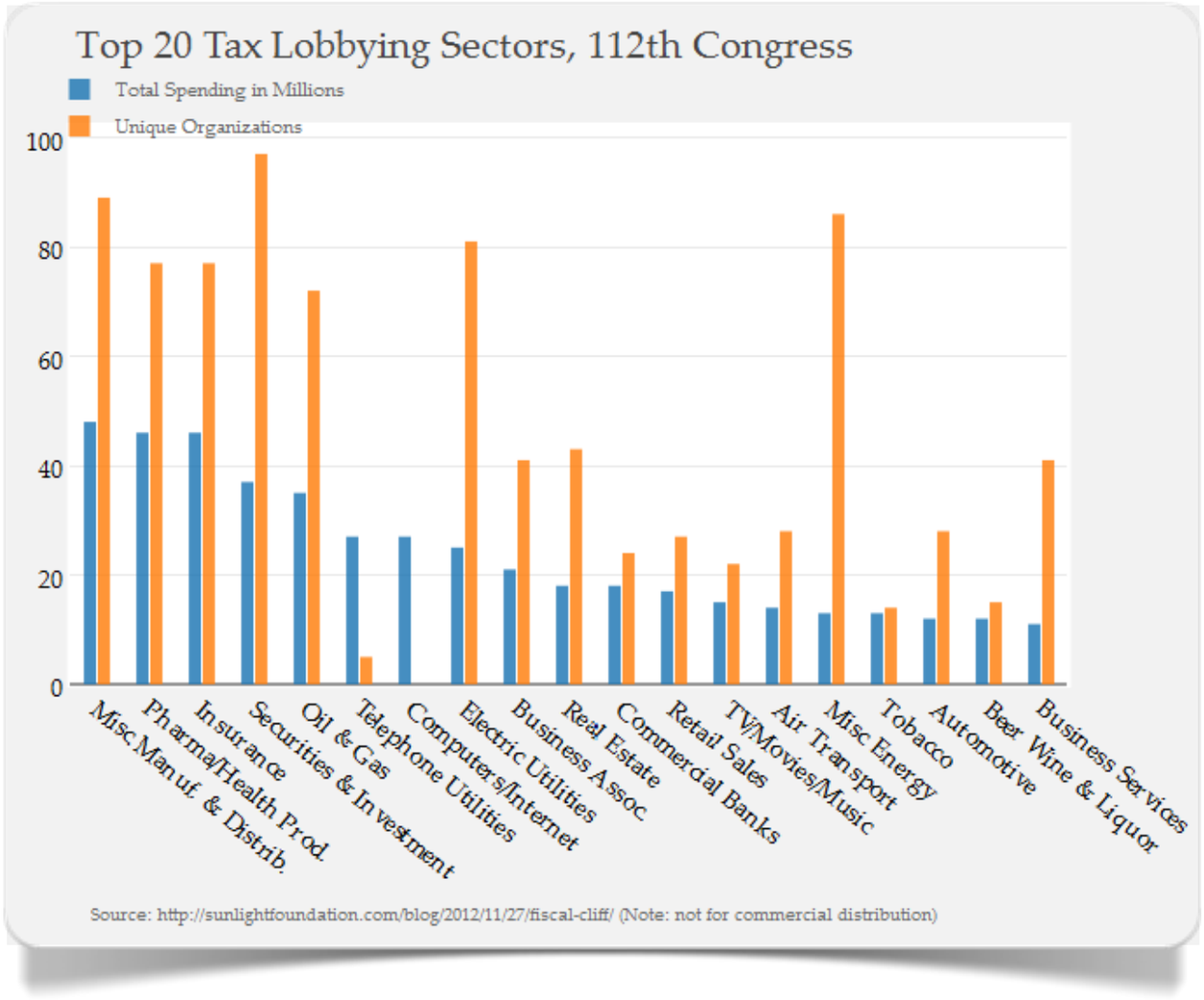


Fig. 3.6 Istogramma categorizzato creato con plotly

Un requisito indispensabile per una corretta rappresentazione del fenomeno che si desidera descrivere è che gli intervalli di variazione delle dimensioni da confrontare abbiano caratteristiche simili per ampiezza e limiti, e che le dimensioni siano in numerosità ridotta.

Attraverso plotly è possibile realizzare istogrammi categorizzati in perfetto stile Microsoft Excel (fig. 3.6).

Per costruire istogrammi categorizzati con Many Eyes è necessario ricorrere al tipo di visualizzazione **bar chart**.

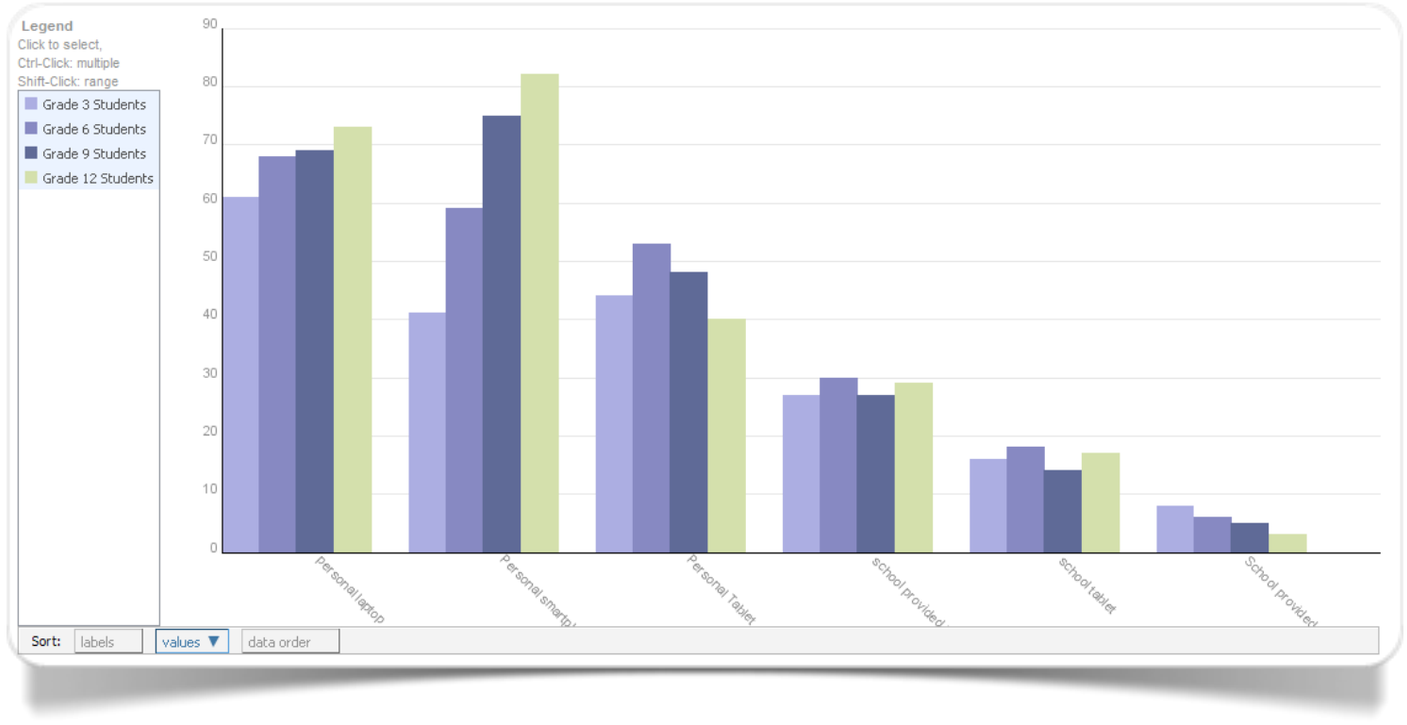


Fig. 3.7 Accesso degli studenti (val. %) agli strumenti mobile personali o forniti dalla scuola per grado scolastico (elab. Many Eyes)

Nella figura 3.7 vediamo un grafico di questo tipo che rappresenta sull'asse delle ordinate la percentuale di studenti che fanno uso a scuola di *mobile devices*: smartphone, tablet o laptop. Nel grafico sono messi a confronto gli studenti che hanno *devices* personali rispetto a quelli che utilizzano quelli forniti dall'istituto scolastico. I dati sono tratti dal rapporto di ricerca dello Speak UP 2012 National Findings: From Chalkboards to Tablets: The Emergence of K-12 Digital Learner (June 2013). Il campione di 365.000 studenti riguarda 8.000 scuole dalle primarie alle secondarie. I "gradi scolastici" e la tipologia di accesso sono rappresentati sull'asse delle ascisse.

Distribuzione di una singola variabile quantitativa (molti dati)

Le **curve di adattamento** si prestano a molti usi. Uno di questi è certamente quello relativo alla rappresentazione “semplificata” di una o più distribuzioni di frequenza. La curva di adattamento consente di evidenziare alcuni aspetti importanti delle singole distribuzioni: tramite di esse è infatti possibile percepire, ad esempio, la presenza di asimmetrie o di sottocampioni provenienti da popolazioni differenti.

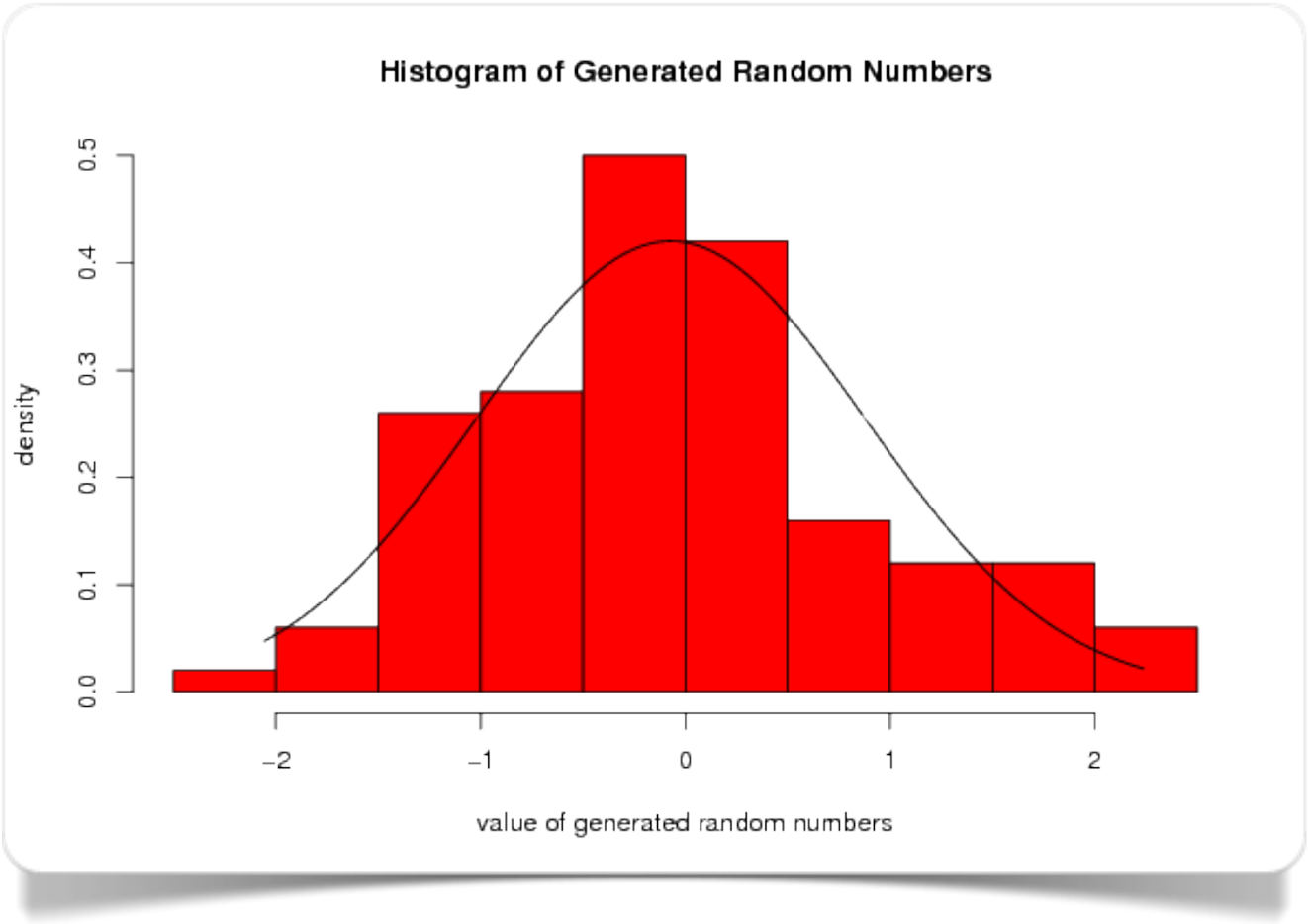


Fig. 3.8 Curva di adattamento creata con Wessa

Questo esempio di curva di adattamento (fig. 3.8; cliccare su *Compute*) è stato realizzato utilizzando Wessa.

In particolare, l’istogramma e la curva sono stati costruiti in seguito a una generazione casuale di dati con distribuzione normale.

Per questo esempio sono state utilizzate le librerie di R *MASS* e *msm*.

Capitolo 4

Distribuzione e Composizione

Grafico a torta

Mappa ad albero

Circle packing

Grafico a barre impilate

Grafico ad aree impilate

Stack Graph

4

Distribuzione e Composizione di una singola variabile categoriale

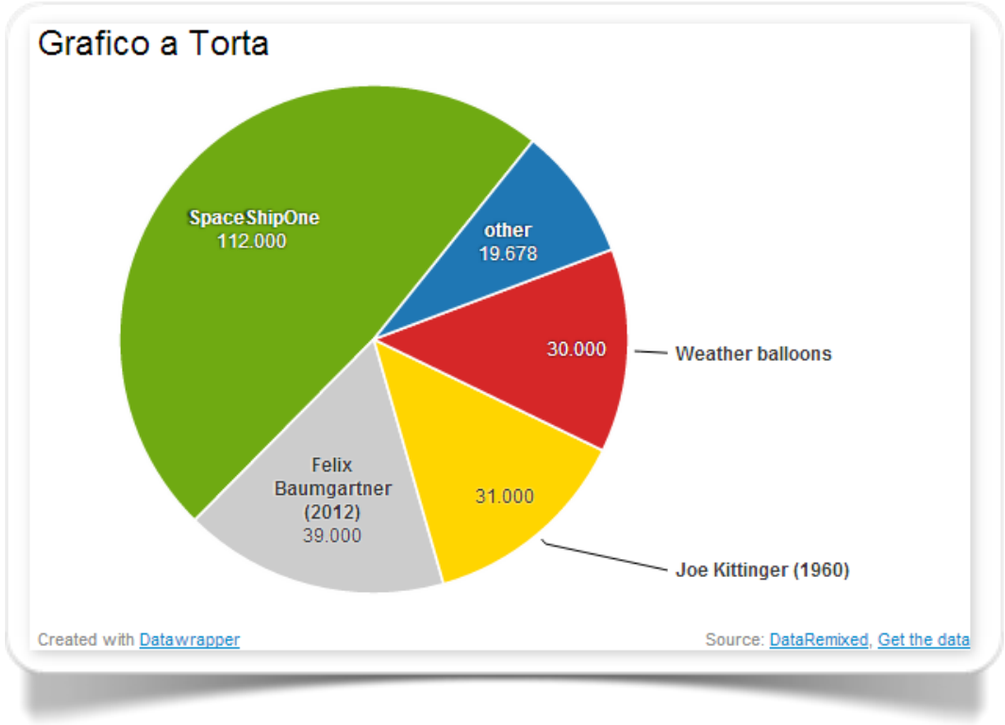


Fig. 4.1 Grafico a torta creato con Datawrapper

Il **grafico a torta** (Playfair, 1801) è una delle rappresentazione grafiche tra le più intuitive: raffigura la distribuzione di frequenza di una **variabile** categoriale (di natura sconnessa o ordinale) quando le categorie disponibili sono di numerosità limitata. Condizione fondamentale e intuitiva perché la rappresentazione sia da considerarsi attendibile è che la somma delle frequenze (percentuali) di tutte le categorie disponibili sia pari al 100%.

Tra le visualizzazioni disponibili in Datawrapper vi sono le pie charts, dotate di numerose opzioni di personalizzazione dei colori, delle etichette e delle grandezze.

Per costruire grafici a torta con Many Eyes è necessario ricorrere al tipo di visualizzazione Pie Chart. Tra le opzioni possibili, la funzione *Slice size* consente di aggiornare i dati sulla base di un'informazione categoriale (ad es., l'anno di riferimento).

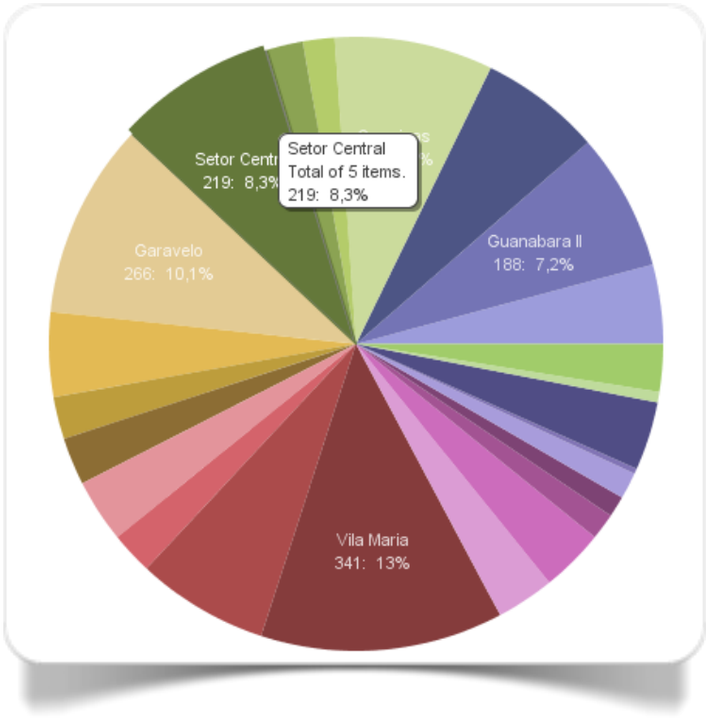
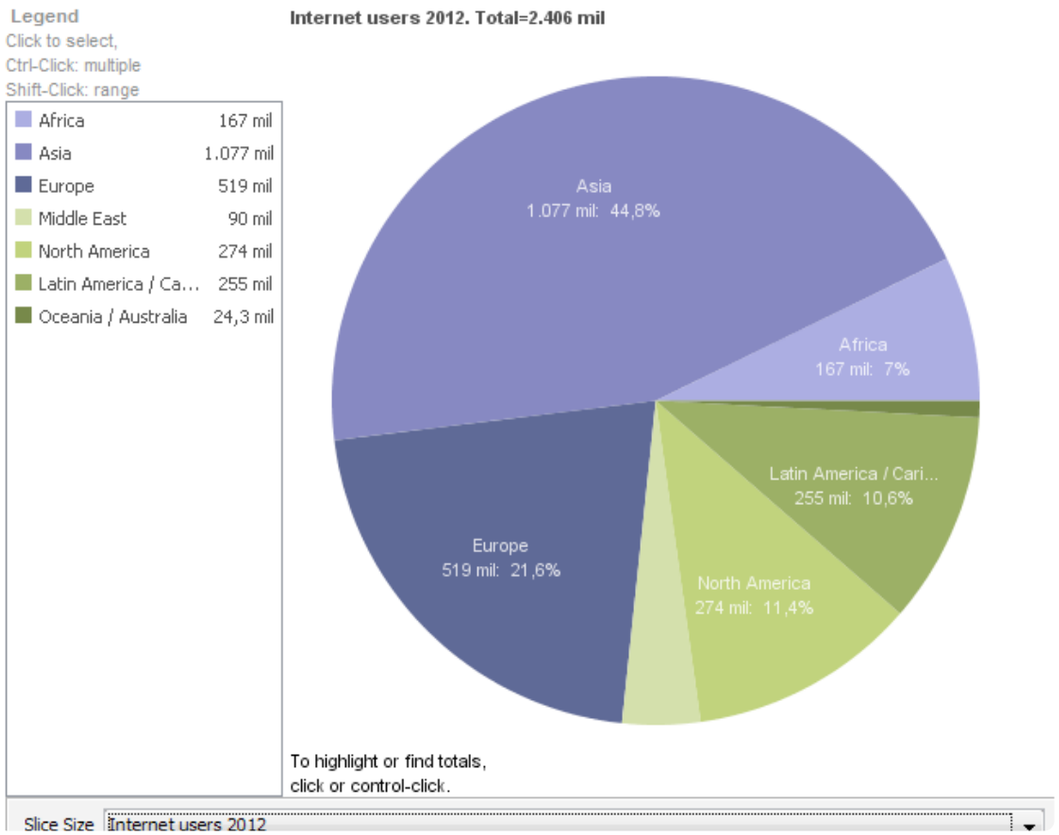


Fig. 4.2 – Grafico a torta creato con Many Eyes

Nella gallery 4.1 possiamo visualizzare il grafico interattivo a torta che permette di confrontare la distribuzione degli utenti internet nel 2012 e nel 2000 (fonte: Internet World Stats). Come si può osservare la situazione è molto cambiata in dodici anni: Nord America ed Europa insieme hanno ceduto il primato all'Asia.



Gallery 4.1 Utenti Internet nel 2012 e nel 2000

Utenti Internet secondo le regioni del mondo nel 2012 (grafico a scorrimento - elab. Many Eyes)

La **mappa ad albero** (Shneiderman, 2009) è una versione alternativa al grafico a torta: la funzione è la medesima (ovvero la rappresentazione di una **distribuzione** di frequenza), si distingue però per la possibilità di rappresentare in maniera gerarchica sotto-distribuzioni. Ogni "quadrante" equivale, cioè, ad una categoria, che a sua volta può rappresentare la somma delle unità appartenenti ad un insieme limitato di sotto-categorie.

Per costruire mappe ad albero con Many Eyes è necessario ricorrere al tipo di visualizzazione Treemap. Questo tipo di grafico dispone di una serie di utili opzioni d'interattività: il principale consiste nella possibilità di modificare le scale dei colori utilizzando un semplice widget disponibile sul fondo del grafico.

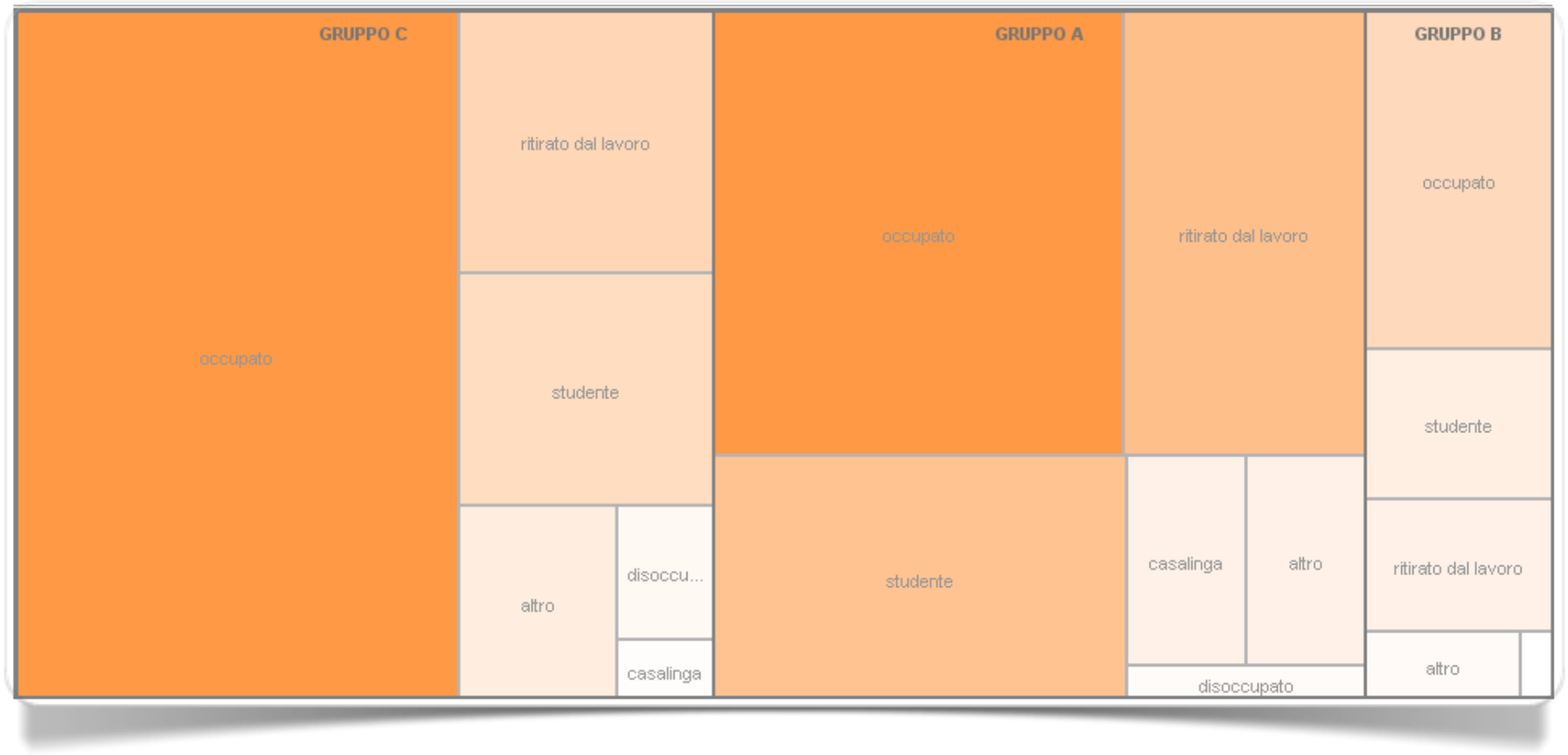


Fig. 4.3 Mappa ad albero creata con Many Eyes

L’esempio della figura 4.4 permette di confrontare tra loro i film che hanno incassato di più nel primo anno di uscita (*#1 Movie*). Il settore più ampio è occupato dal film che ha incassato di più in assoluto: *Titanic*. La colorazione in ocra ci dice che la differenza di incasso nello stesso anno rispetto al film che ha vinto l’Oscar (*Best Picture*) è pari a zero, perché Titanic nel 1997 ha vinto anche l’Academy Award. Lo stessa cosa possiamo dire per *The lord of the ring* (2003) e per *Rain Man* (1988). Invece tra l’incasso di *The Lion King* e *Forrest Gamp*, il vincitore del premio Oscar in quell’anno (1994), la differenza è notevole (il 28,7% in meno). La colorazione in azzurro intenso denota una differenza maggiore che diminuisce avvicinandosi al bianco. L’intensità del colore è selezionabile nel grafico interattivo con la barra in basso a destra.

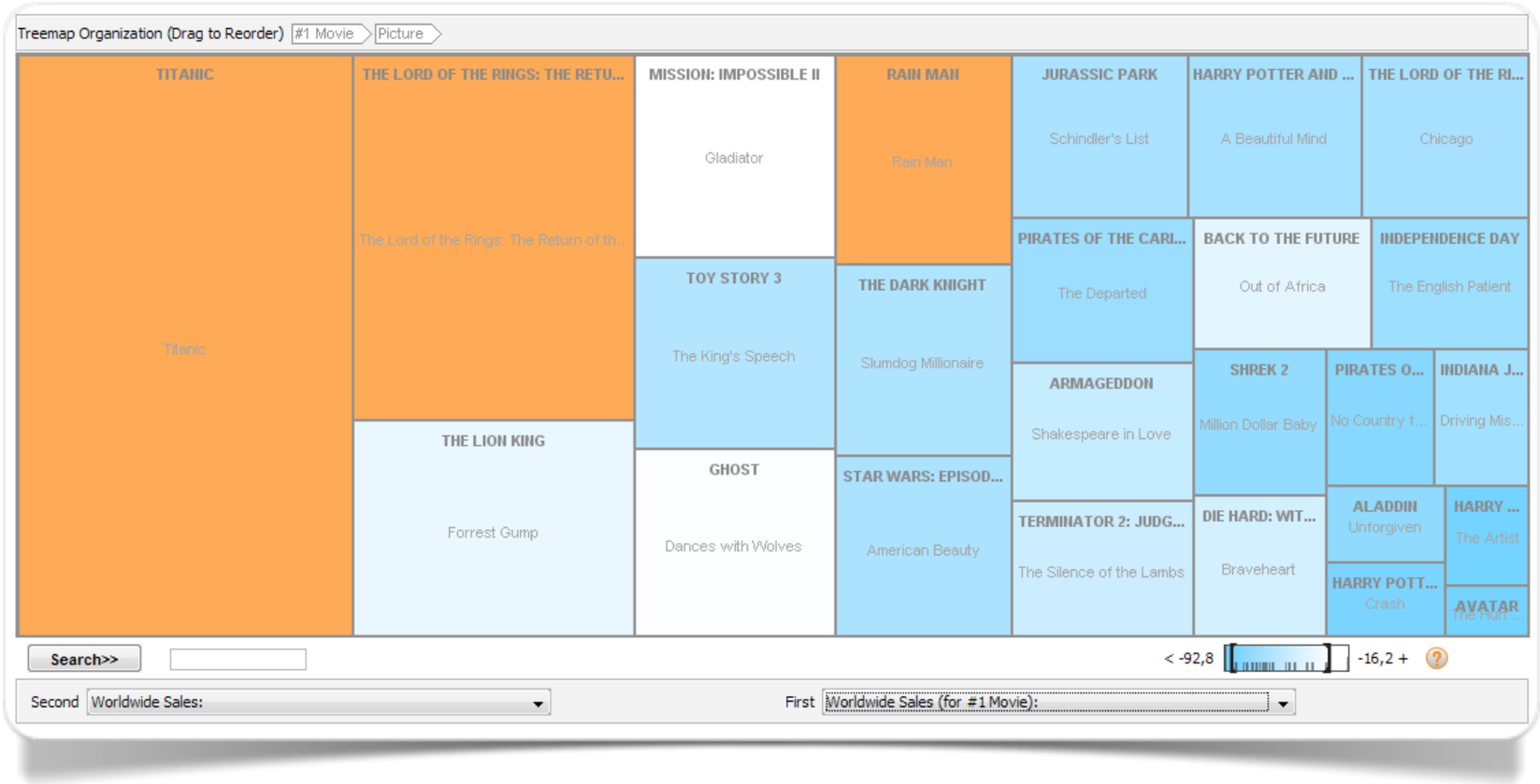


Fig. 4.4 Incasso in dollari dei film di maggior successo - #1 Movie - dal 1978 al 2011 (elab. Many Eyes)

Il grafico **treemap** di Many Eyes permette molte altre opzioni. Oltre alla visualizzazione delle informazioni muovendo il mouse sull'area di interesse, è possibile, visualizzare i dati relativi ai migliori film (*Best Picture*) mettendo in evidenza i vincitori degli Oscar rispetto ai film di maggiore incasso. Il cambiamento di prospettiva si ottiene cliccando e trascinando in prima posizione la linguetta in alto a sinistra sulla quale è riporta la "categoria" in analisi: *Picture* passa in primo piano e *#1 Movie* in secondo piano (fig. 4.5).

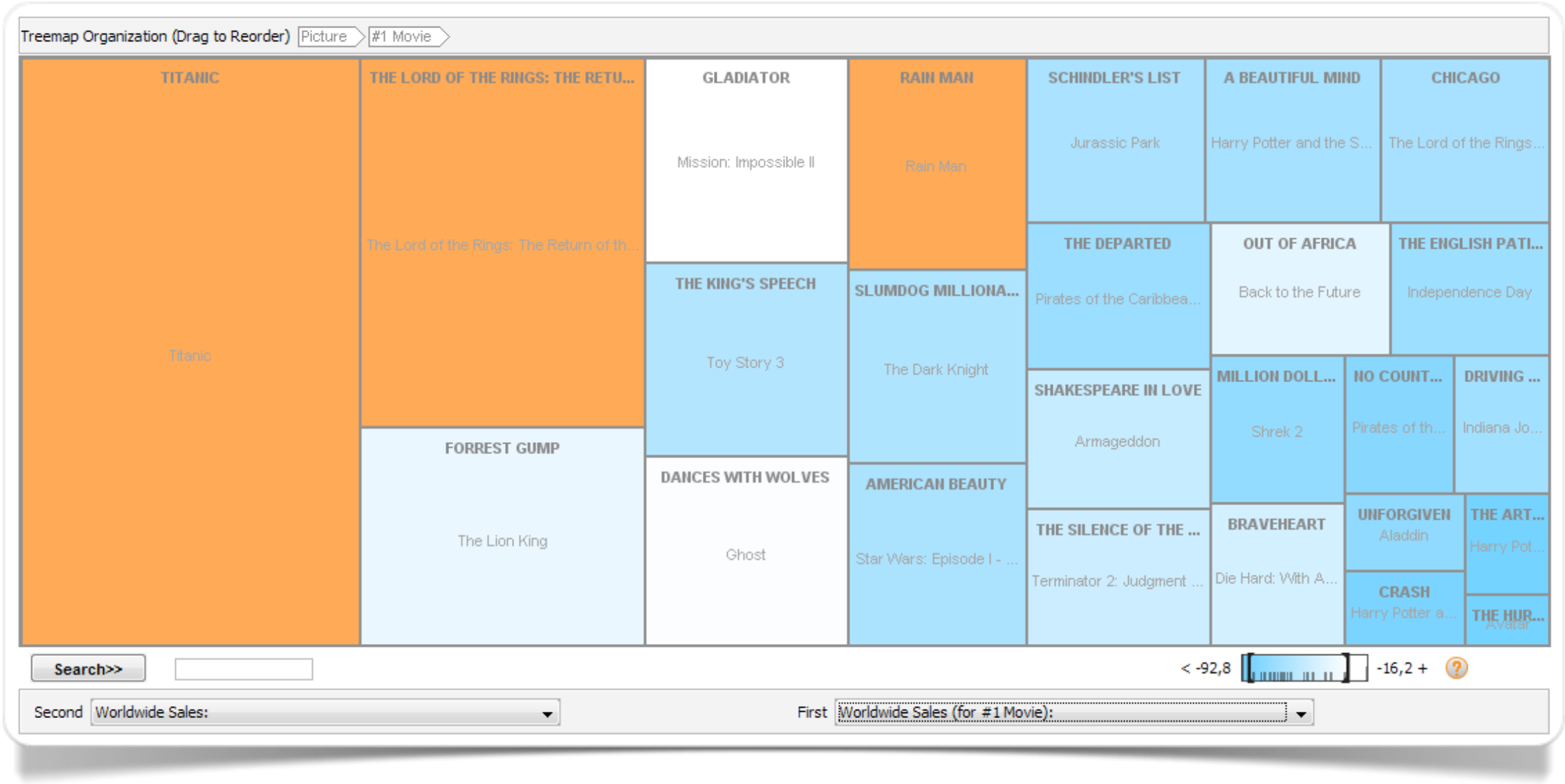


Fig. 4.5 Incasso in dollari dei film di maggior successo - Best Picture - dal 1978 al 2011 (elab. Many Eyes)

Fig. 4.6 Mappa ad albero realizzata con Raw

Attraverso Raw è possibile realizzare mappe ad albero attraverso l'impiego di un'unica semplicissima interfaccia (fig. 4.8). A differenza di molti altri tool simili, Raw non salva i dati su alcun database.

Il **circle packing** è un particolare studio geometrico che si basa sull'occupazione di aree o superfici con una serie di cerchi di raggio uguale o variabile con il principale vincolo di evitare sovrapposizioni. Da relativamente poco tempo, questo tipo di studio viene utilizzato per creare raffigurazioni statistiche simili ai treemap per scopo: in questo caso l'organizzazione del grafico è tale che sia le variabili categoriali relative alle gerarchie superiori che quelle relative alle gerarchie inferiori rispettino il principio alla base del circle packing.

Attraverso Raw è possibile realizzare circle packing in cerchi (si veda Wikipedia). La loro realizzazione in Raw è possibile selezionando tante variabili categoriali tante quante sono le gerarchie che si desiderano organizzare in cerchi. La dimensione di ogni cerchio è definita non dalla frequenza di ogni combinazione di gerarchie bensì dalla specificazione di una **variabile** numerica.

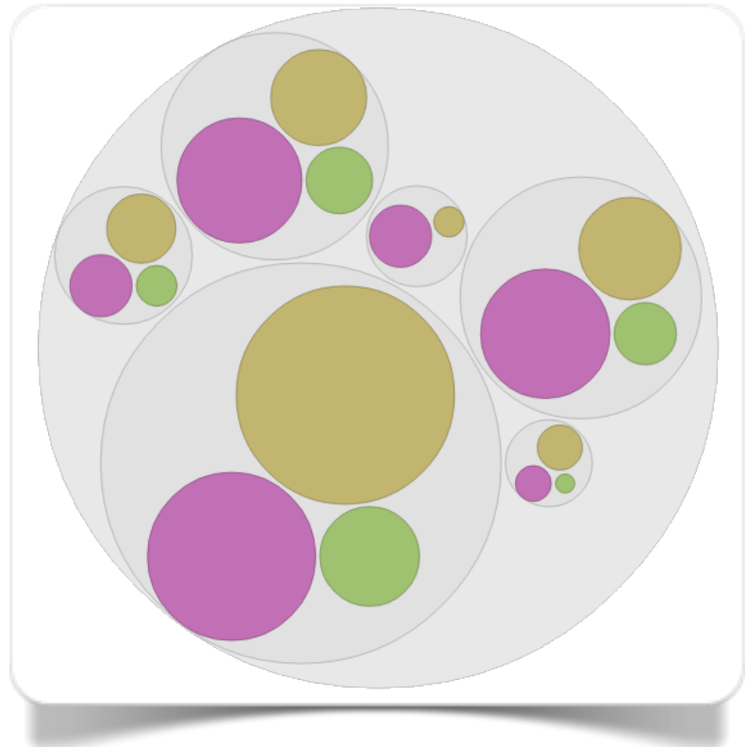


Fig. 4.7 Circle packing creato con Raw

D&C di una singola variabile categoriale segmentata (valori assoluti)

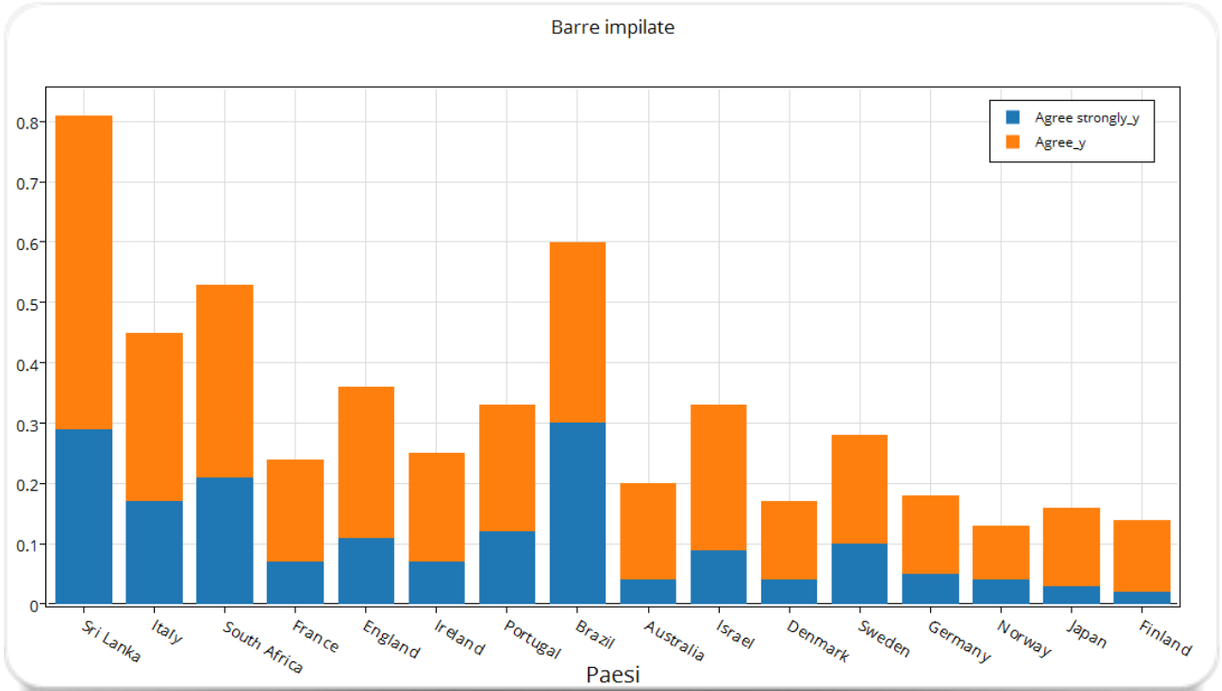


Fig. 4.8 Grafico a barre impilate realizzato con plotly

plotly consente di costruire grafici a barre impilate ricorrendo al tipo di visualizzazione Bar Charts (fig. 4.8).

Anche Datawrapper offre la possibilità di creare grafici a barre impilate come nell'esempio che potete trovare qui di fianco (fig. 4.9).

Il **grafico a barre impilate** è lo strumento ideale per visualizzare la **distribuzione** delle occorrenze di ognuna delle categorie di una specifica **variabile** categoriale (qualitativa) lungo i diversi livelli di una seconda variabile categoriale (X).

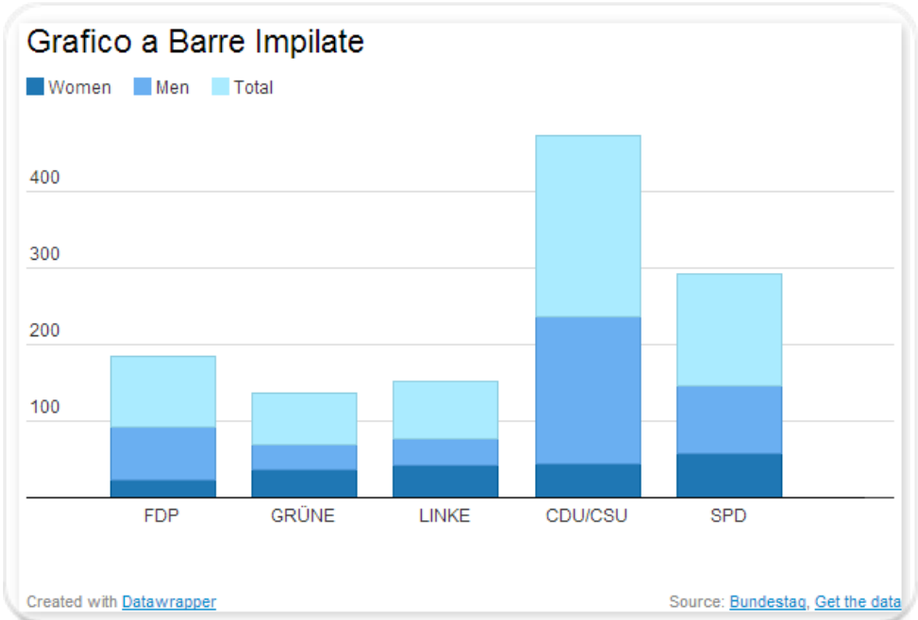


Fig. 4.9 Grafico a barre impilate realizzato con Datawrapper

Il **grafico ad aree impilate** è una versione corretta del grafico a barre impilate. La differenza rispetto a quest'ultimo grafico è la tipica rappresentazione "continua" dei valori lungo l'asse orizzontale. Le aree impilate trovano la loro applicazione ideale quando sull'asse orizzontale è riportata una dimensione temporale: il caratteristico andamento "continuo" delle curve che delineano le aree consentono di rivelare al meglio le eventuali tendenze ed evoluzioni nel tempo.

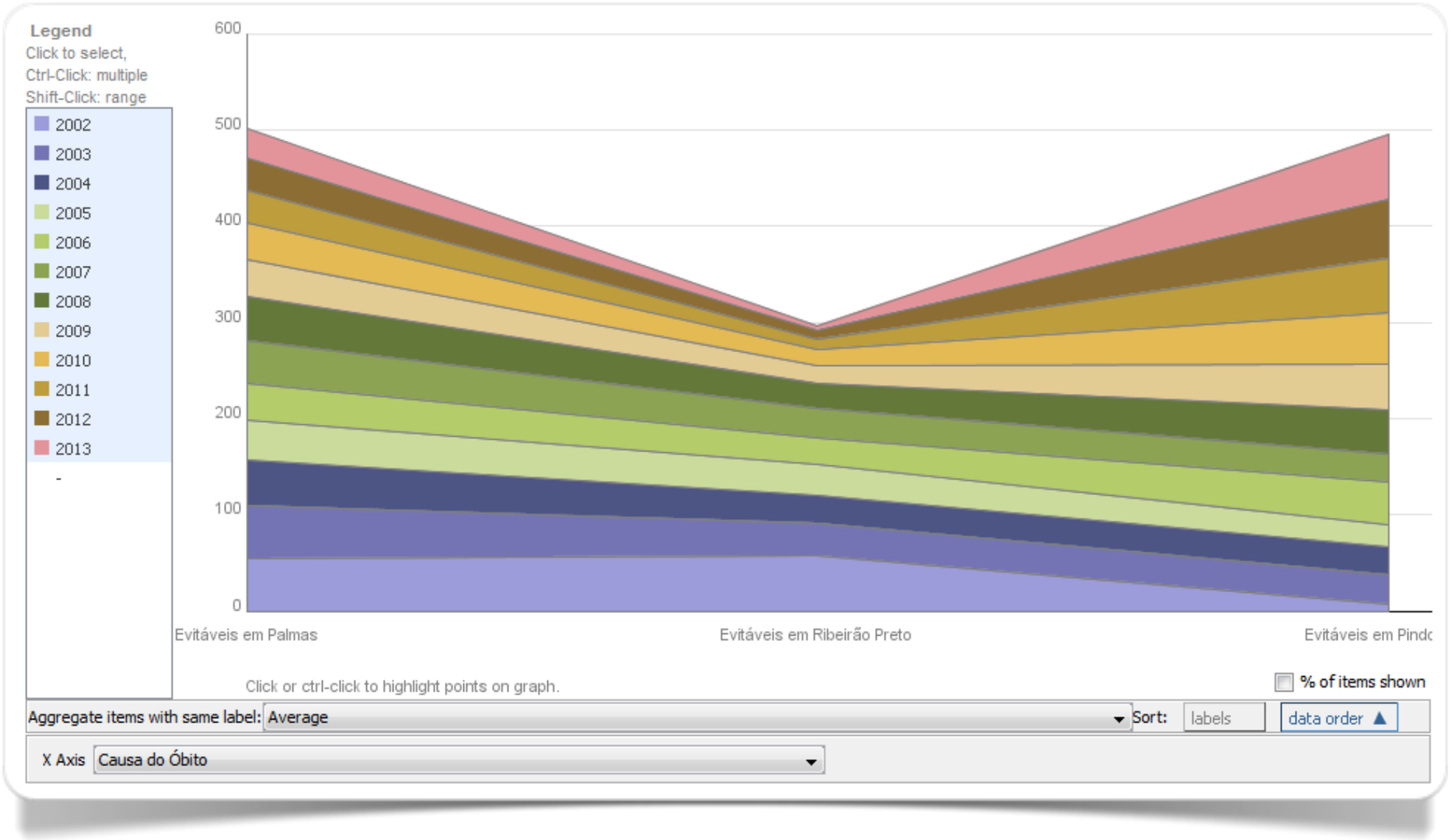


Fig. 4.10 Grafico ad aree impilate realizzato con Many Eyes

Gli Stack Graphs sono tra i grafici esteticamente più efficaci tra quelli disponibili in Many Eyes. Nella figura 4.11 vediamo la distribuzione della popolazione per contea in una serie temporale dal 1841 al 2011 (fonte: Central Statistic Office Ireland). Sull'asse delle ordinate è rappresentato l'ammontare della popolazione e sull'asse delle ascisse sono rappresentati gli anni di rilevazione. Nel grafico interattivo è selezionabile la visualizzazione della singola contea oppure le distribuzioni percentuali della popolazione nelle contee per ciascun anno. Questo permette, per esempio, di osservare con maggiore evidenza l'effetto dell'urbanizzazione con l'accrescimento delle contee di Dublino e Cork a discapito delle altre.

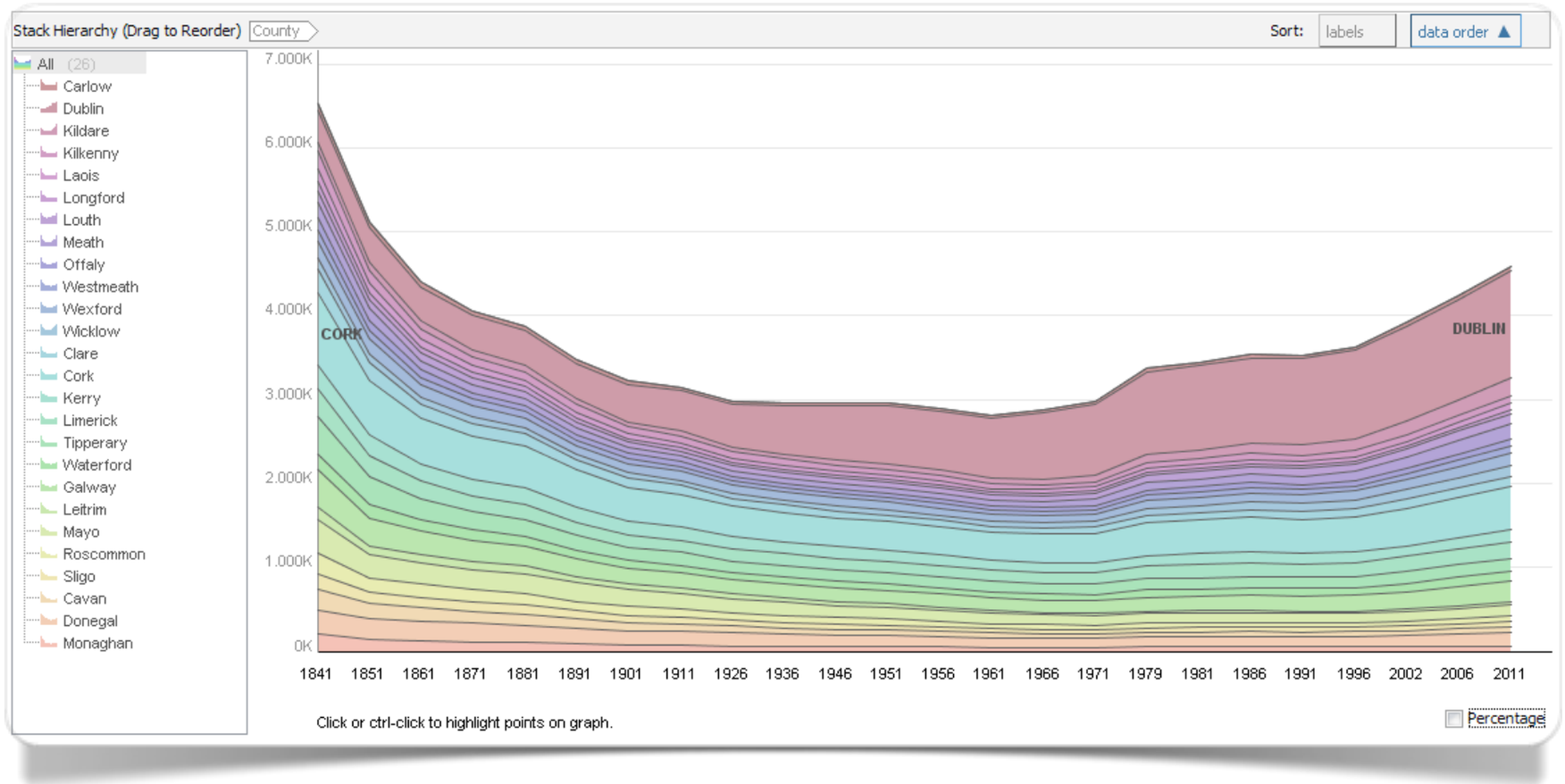


Fig. 4.11 Popolazione residente in Irlanda secondo la contea dal 1841 al 2011 (elab. Many Eyes)

D&C di una singola variabile categoriale segmentata (val. %)

Il grafico a **barre impilate** è lo strumento ideale per visualizzare la **distribuzione** di frequenza (delle percentuali) di ognuna delle categorie di una specifica **variabile** categoriale (qualitativa) lungo i diversi livelli di una seconda variabile categoriale (X).

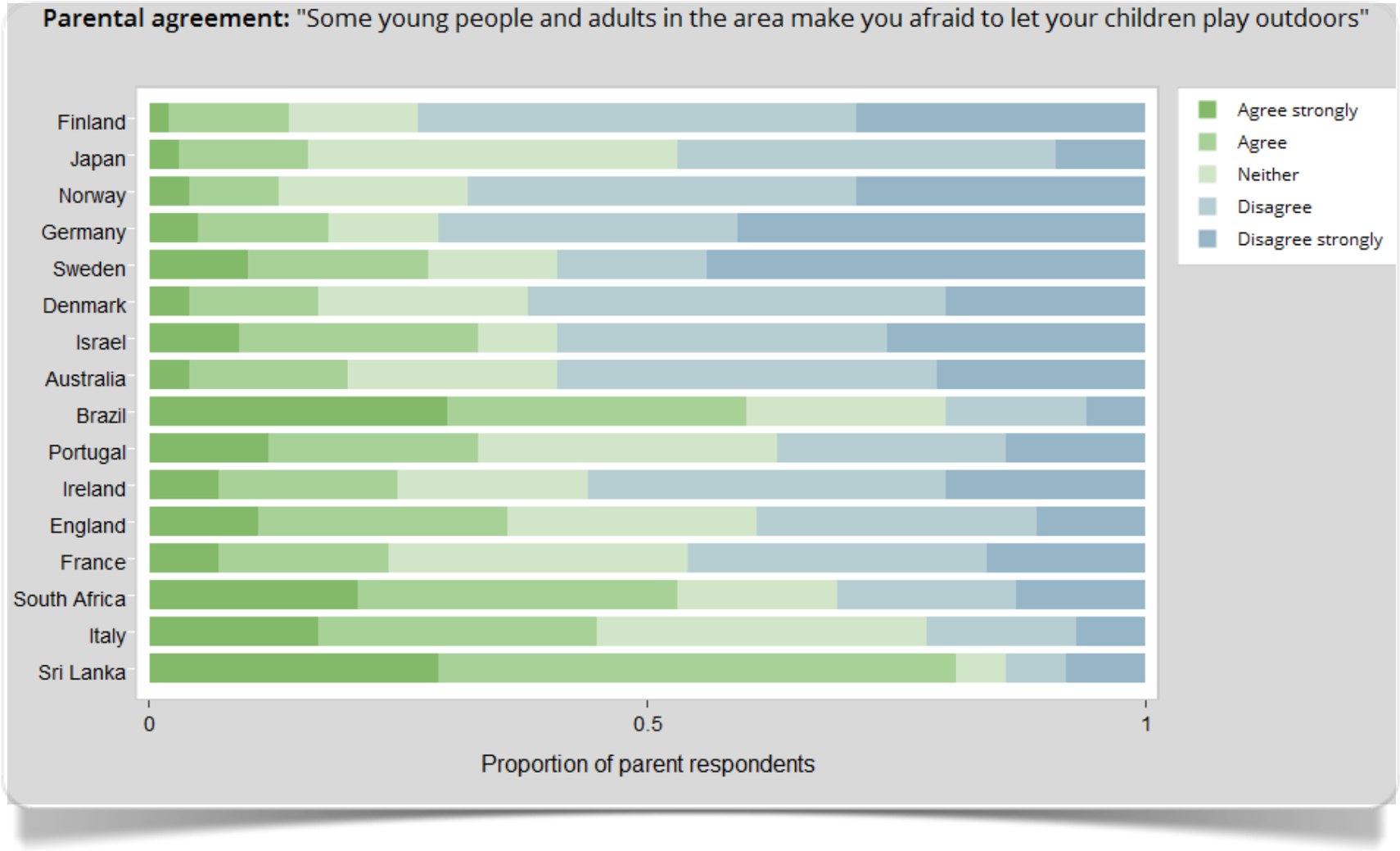


Fig. 4.12 Grafico a barre impilate realizzato con plotly

La funzionalità di creazione di grafico *Bar charts* disponibile in plotly consente di costruire facilmente grafici a barre impilate. Ciò è possibile selezionando la variabile categoriale come asse X o Y del grafico (in fig. 4.13 la variabile paesi è stata selezionata come asse Y) e facendo una selezione multipla di variabili numeriche (percentuali) a somma 100.

Many Eyes consente di costruire grafici a barre impilate ricorrendo al tipo di visualizzazione Matrix Chart. Se si sceglie l'opzione della rappresentazione a barre (*Bars*) è possibile utilizzare l'opzione *Same size* per riportare su grafico le dimensioni delle diverse categorie su base proporzionale (percentuale).

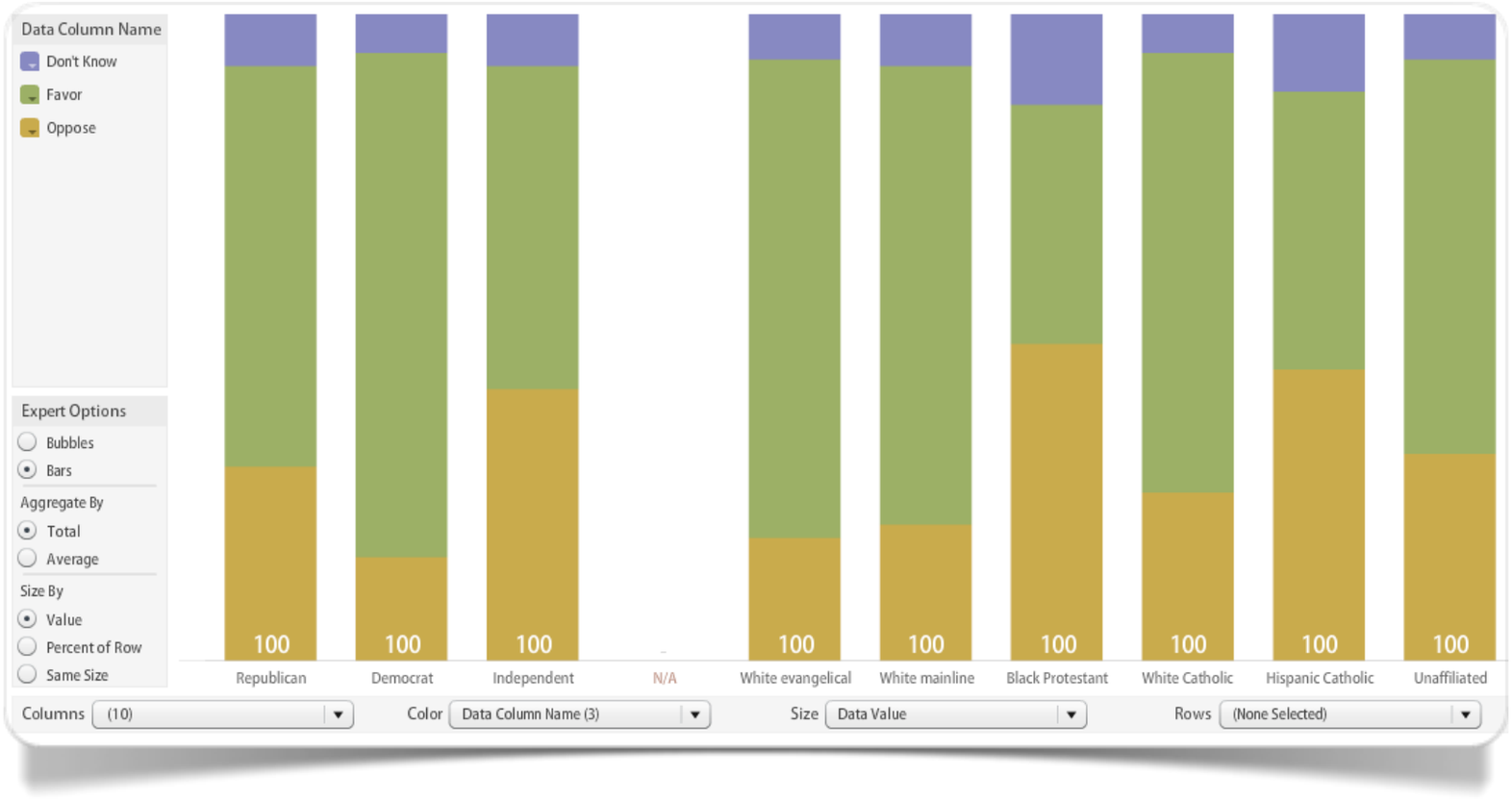


Fig. 4.13 Opinioni sulla pena di morte negli Stati Uniti per orientamento politico e confessione religiosa (val. % - elab.Many Eyes)

Nella figura 4.13 sono rappresentate le opinioni di un campione di cittadini degli Stati Uniti sulla pena di morte (fonte: Pew Research Center Survey 2010). I valori in percentuale delle modalità di risposta (favorevole, contrario, non so) sono rappresentate sull'asse delle ordinate. Le categorie sono messe a confronto sull'asse delle ascisse, in due raggruppamenti principali: orientamento politico e confessione religiosa, con una sub-categoria per bianchi, neri e ispanici. Come possiamo osservare sono ancora numerosi i favorevoli alla pena di morte. I più contrari si trovano tra i neri protestanti e gli ispanici cattolici. L'appartenenza etnico-culturale è più forte del riferimento religioso. Nel grafico interattivo si può selezionare una rappresentazione con grafici a torta in sostituzione delle barre, ma la visualizzazione è meno efficace.

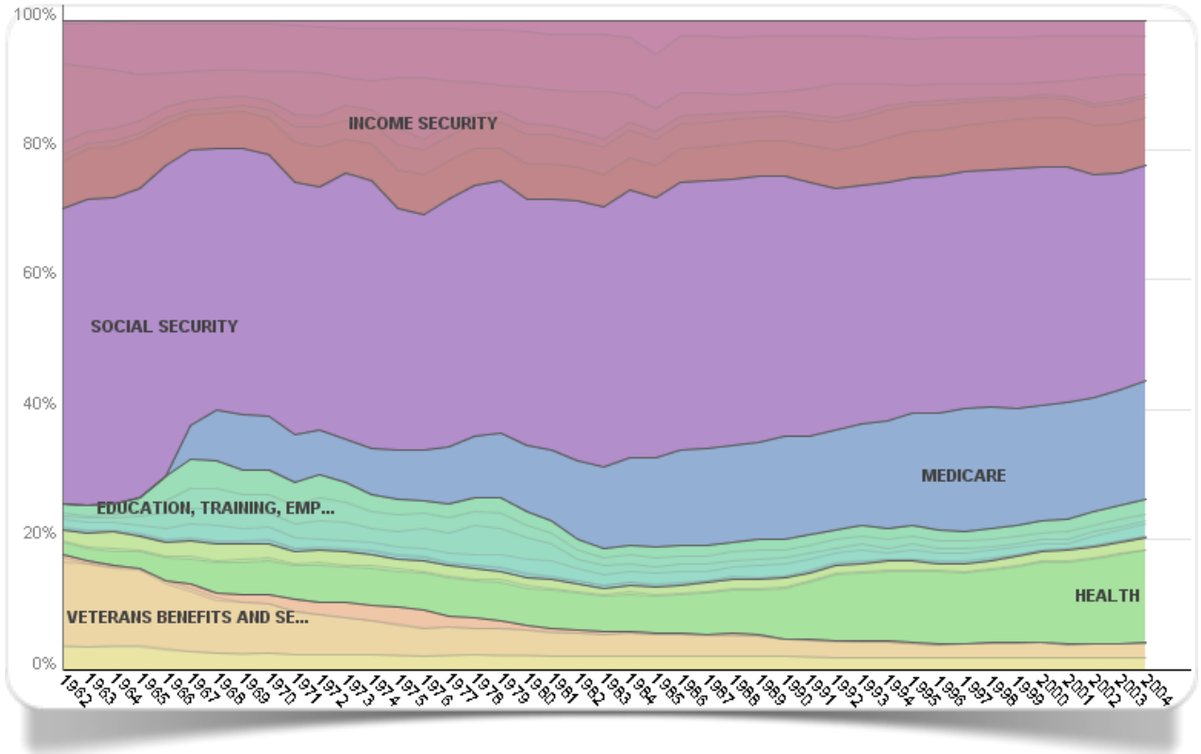


Fig. 4.14 Grafico ad aree impilate realizzato con Many Eyes

Il grafico ad **aree impilate** può essere utilizzato quale versione corretta del grafico a barre impilate per percentuali.

Gli **stack graphs** sono tra i grafici esteticamente più efficaci tra quelli messi a disposizione da Many Eyes. Attraverso l'opzione *Percentage* è possibile visualizzare la distribuzione percentuale delle frequenze delle diverse categorie.

Capitolo 5

Relazione e Distribuzione

Istogramma a blocchi

5

Relazione e Distribuzione
variabile categoriale vs una singola variabile numerica

L'**istogramma a blocchi** è un particolare grafico il cui principale obiettivo è rappresentare la distribuzione delle classi di una variabile categoriale rispetto ai valori di una variabile numerica. Solitamente il suo utilizzo prevede la rappresentazione della variabile numerica sull'asse X e la determinazione di suoi sotto-intervalli, in corrispondenza dei quali "impilare" una serie di blocchi ognuno corrispondente ad una specifica classe della variabile categoriale. L'ordinamento verticale dei blocchi (classi) è tale, per cui le classi a cui è associato un valore di X più elevato – nel rispettivo sotto-intervallo di valori - saranno più in alto sull'asse Y rispetto a quelle a cui è associato un valore di X più basso.

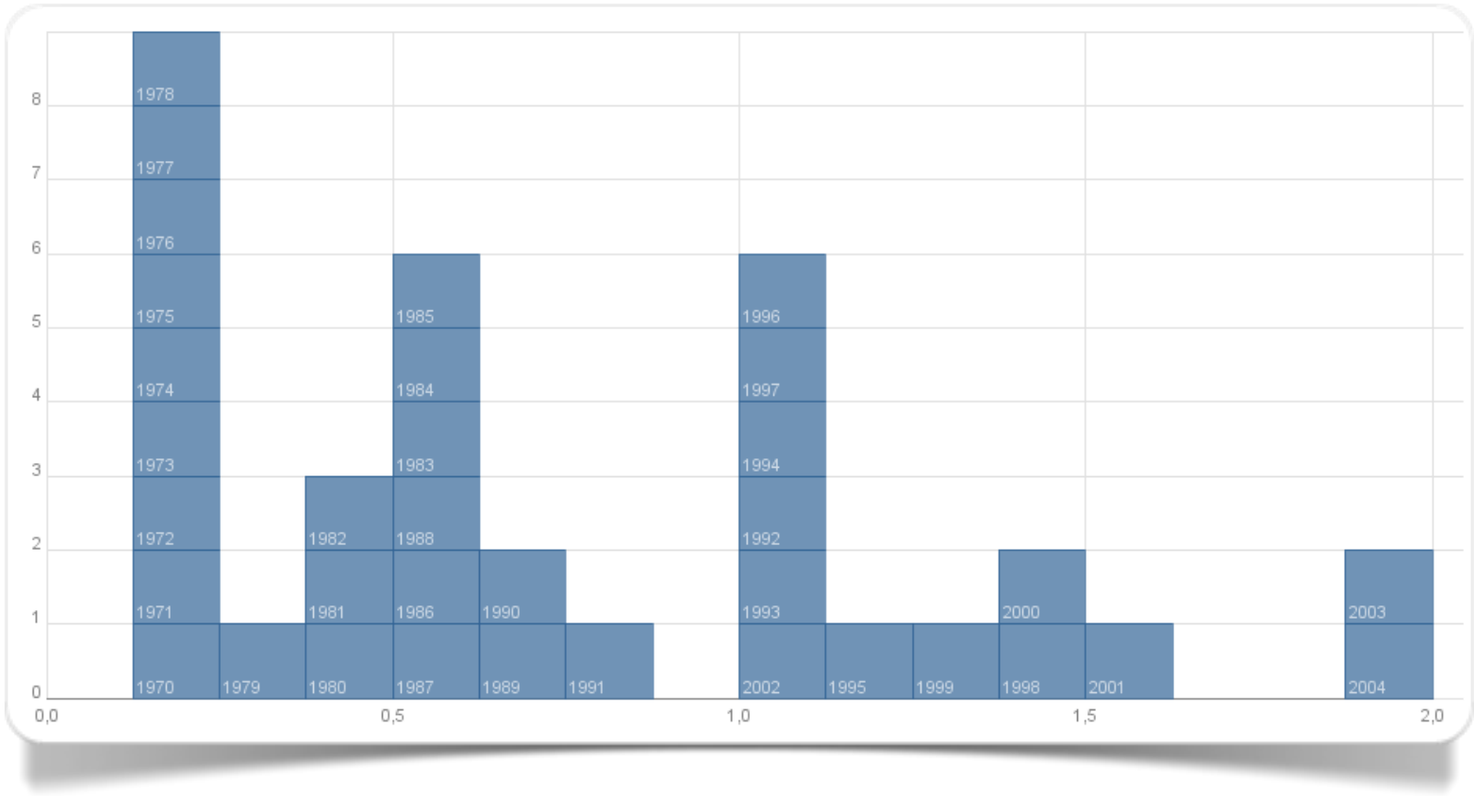


Fig. 5.1 Grafico ad aree impilate realizzato con Many Eyes

Tra gli strumenti di creazione su web di grafici statistici Many Eyes risulta essere attualmente l'unico che permette di realizzare istogrammi a blocchi. Nell'esempio 5.2 possiamo mettere a confronto i paesi europei secondo alcuni parametri economici. Il grafico interattivo on line permette di selezionare le variabili di interesse e di visualizzare il riordinamento dei paesi secondo la selezione.

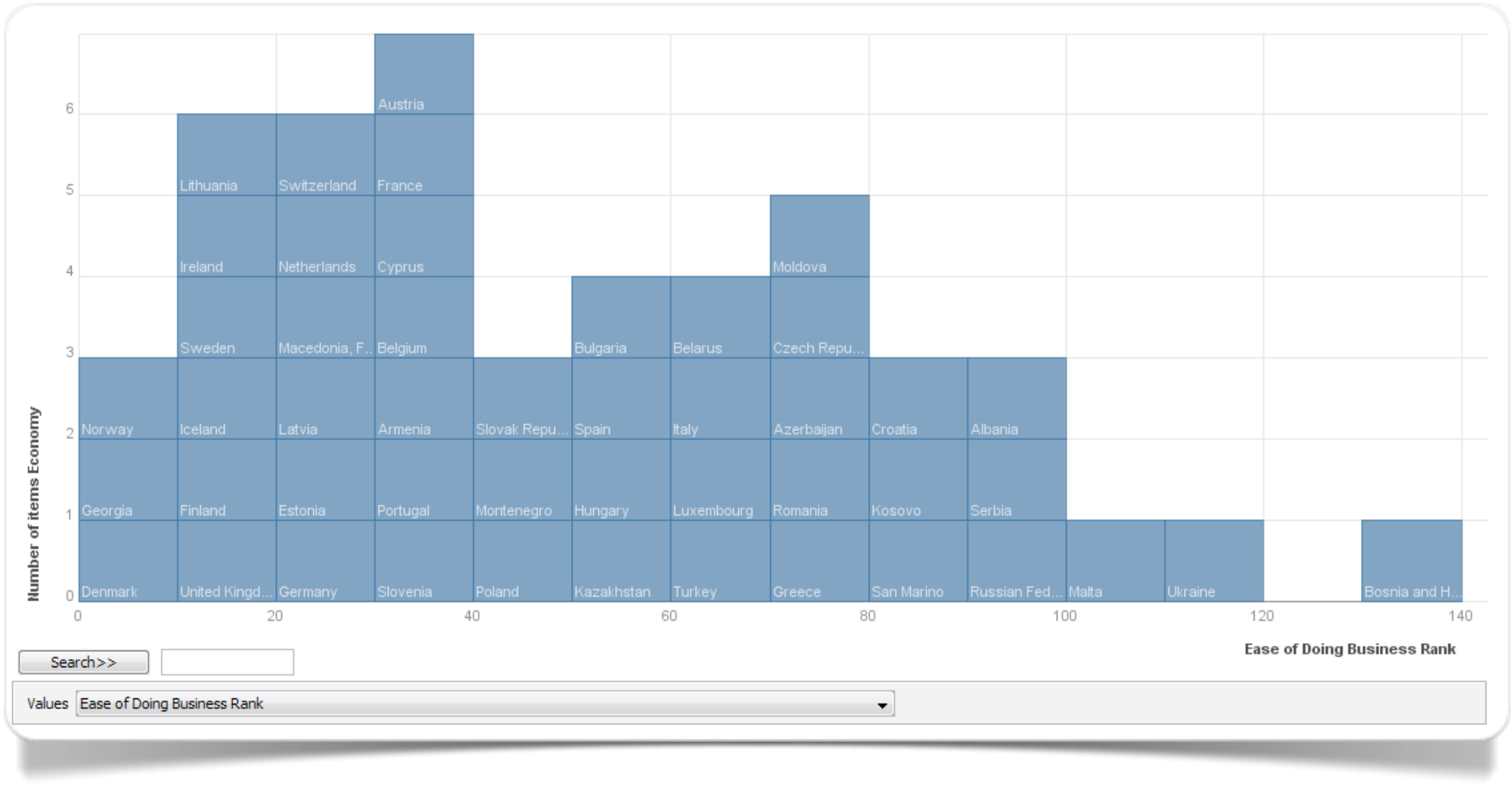


Fig. 5.2 Istogramma a blocchi con la classifica delle economie europee (elab. Many Eyes)

Capitolo 6

Relazione e Composizione

Coordinate parallele

6

Relazione e Composizione tra molte variabili quantitative

Le **coordinate parallele** (d'Ocagne, 1885; Inselberg, 1985) è uno strumento utilizzato per visualizzare e analizzare dati multivariati. Nell'area grafica vengono riportate tante linee verticali (assi o dimensioni) parallele equidistanti l'una dall'altra quante sono le variabili numeriche da rappresentare.

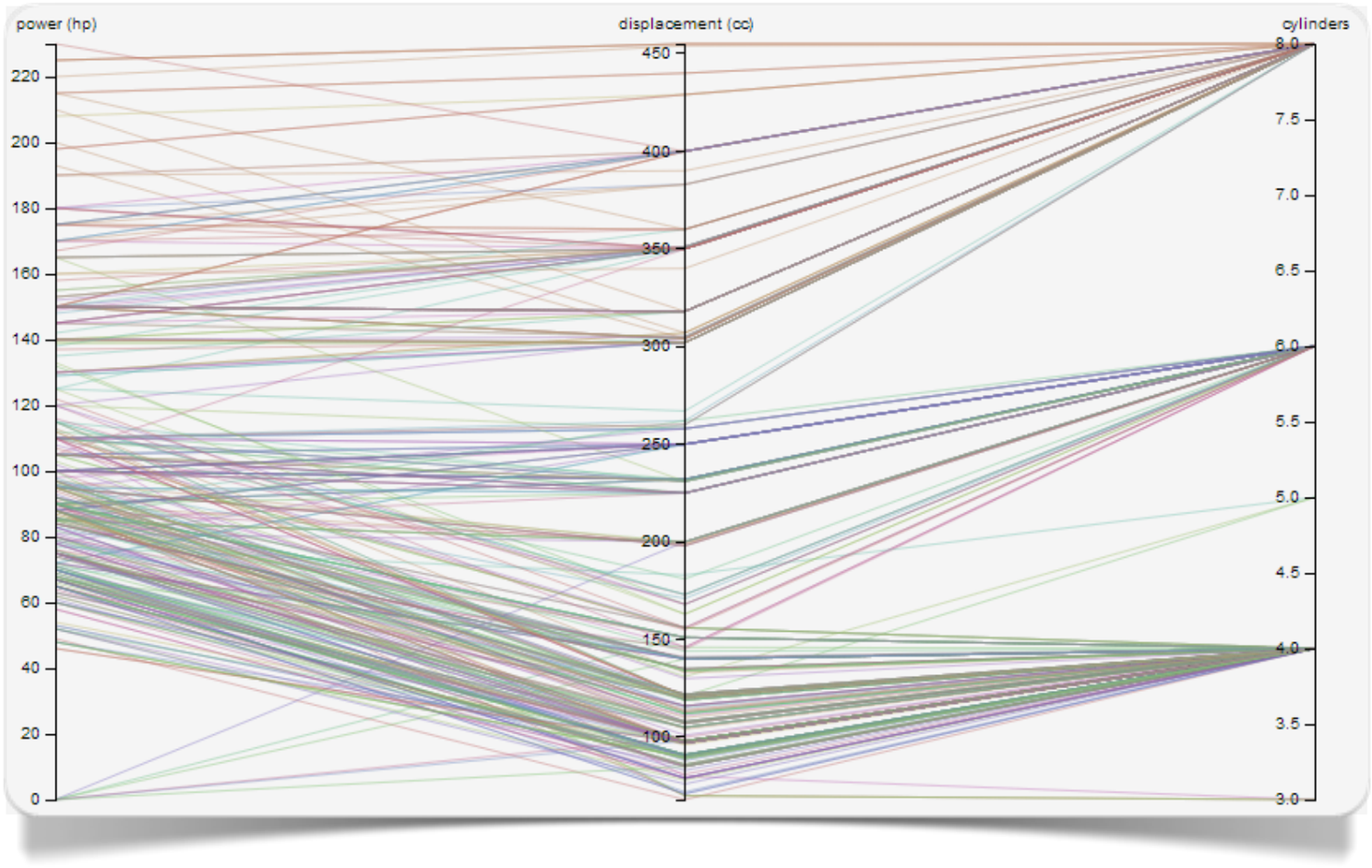


Fig. 6.1 Coordinate Parallele realizzate con Raw

Le unità statistiche vengono rappresentate da una linea spezzata i cui vertici si collocano esattamente in corrispondenza di ogni linea verticale. La posizione del vertice sull'i-esimo asse corrisponde all'i-esima coordinata del punto (vedi Wikipedia).

Raw consente di selezionare un numero n di **variabili** numeriche da utilizzare come linee verticali (assi o dimensioni). Oltre a ciò è possibile selezionare una variabile categoriale in modo da identificare ad esempio con uno stesso colore le unità appartenenti ad una stessa classe.

L'**alluvial diagram** è un particolare tipo di **diagramma di flusso**, in cui le variabili qualitative vengono organizzate per linee parallele e per blocchi (cluster di nodi). Al passaggio da una variabile qualitativa all'altra si osservano diramazioni o accorpamenti dei diversi flussi sulla base dell'appartenenza contemporanea di ogni singola unità statistica alle diverse classi delle variabili qualitative coinvolte. Al flusso di ogni cluster è possibile assegnare uno "spessore" attraverso una o più variabile quantitativa di dimensione.

Raw consente di selezionare un numero di variabili qualitative da utilizzare come step dell'alluvial diagram. Opzionalmente è anche possibile utilizzare variabili numeriche per la creazione di nuovi step.

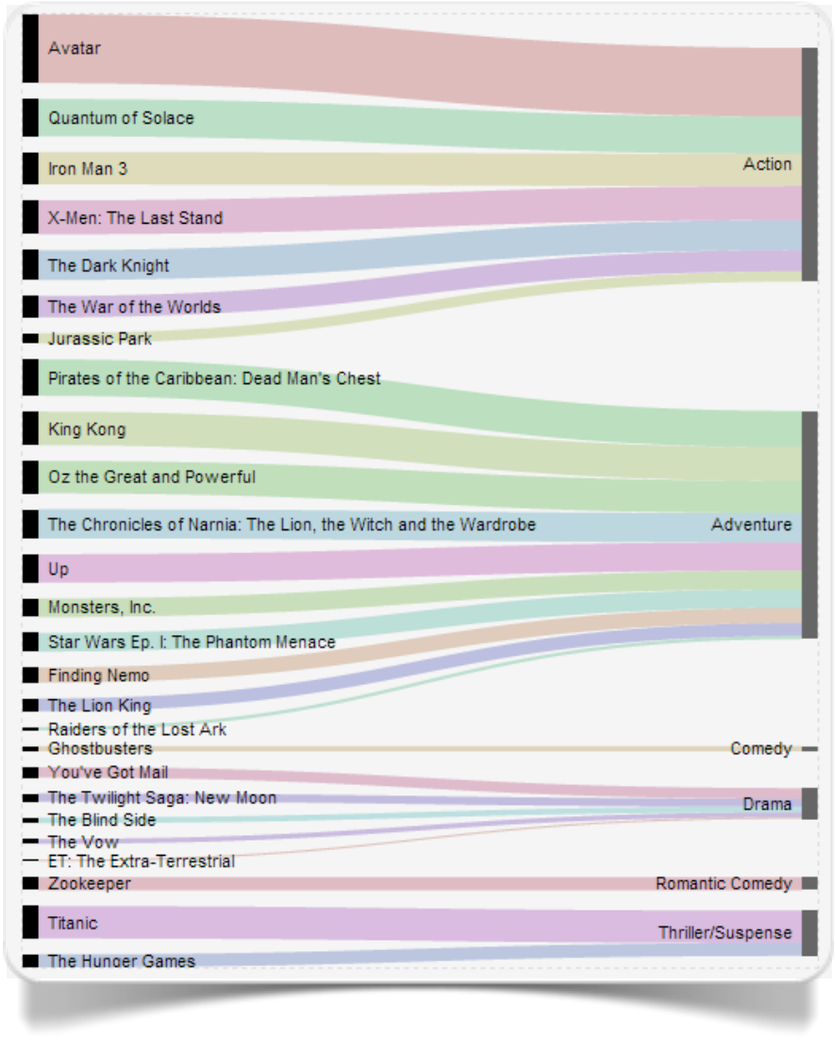


Fig. 6.2 Alluvial diagram realizzato con Raw

Capitolo 7

Confronto e Distribuzione

Box and whiskers plot

Bagplot

7

Confronto e Distribuzione tra misure di posizione e misure di dispersione

Noto come **Box-plot**, o meglio ancora come **box and whiskers plot** (*diagramma a scatole e baffi*; Tukey, 1977), questo tipo di grafico è principalmente utilizzato in statistica per confrontare le posizioni (media, mediana, ecc.) e le misure di dispersione (deviazione standard, intervallo interquartile, ecc.) lungo diversi gruppi di unità appartenenti ad una medesima **variabile**. L'ulteriore vantaggio di questa rappresentazione risiede nella possibilità d'interpretare la natura distributiva dei dati all'interno di ogni gruppo. I baffi, in special modo, possono consentire di evidenziare una maggiore o una minore dispersione al di sotto o al di sopra della rispettiva misura di posizione.

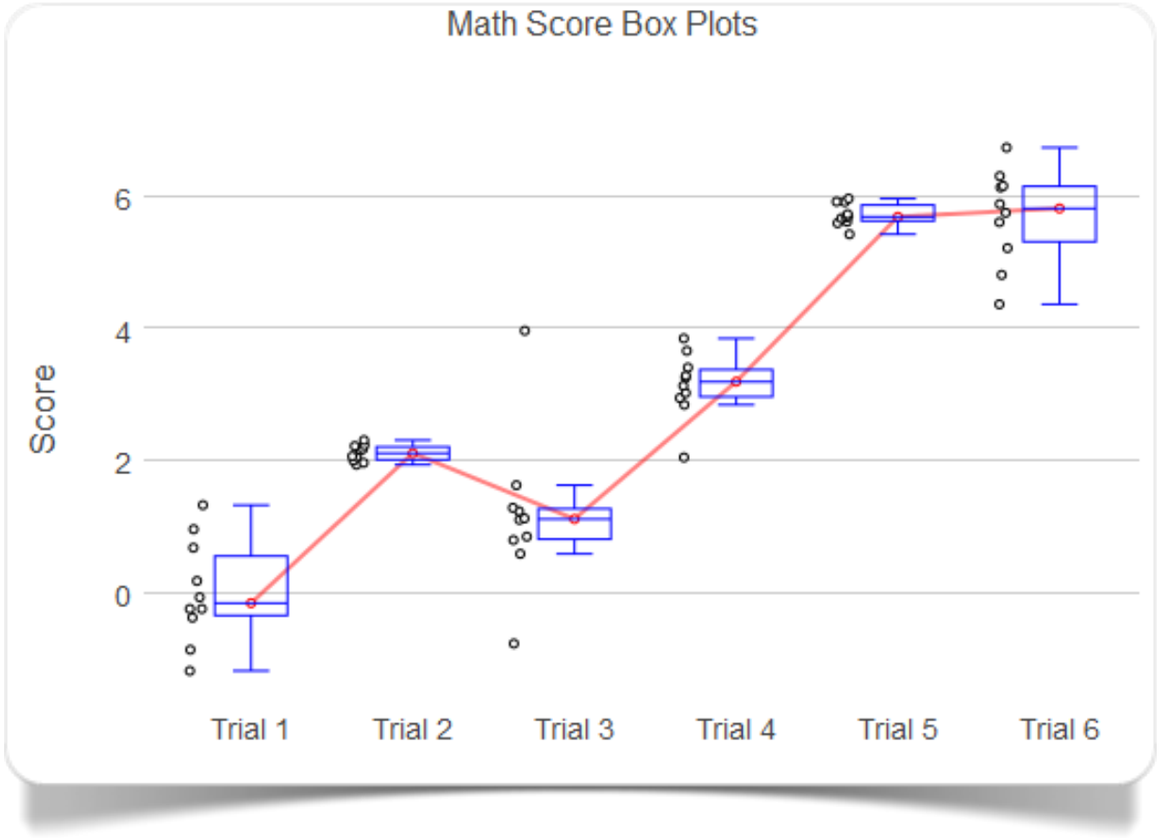


Fig. 7.1 Box and Whiskers Plot realizzato con plotly

I box and whiskers plot possono essere creati con plotly (come in figura 7.1). Il programma richiede all'utente di specificare soltanto una colonna di valori numerici per ogni box and whiskers plot, dopodiché provvede in automatico a calcolarne le rispettive misure di tendenza centrale e di variazione.

Confronto e Distribuzione tra due variabili quantitative

Il **bagplot** (Rousseeuw e al., 1999) è la rappresentazione bidimensionale del box-plot. Nel bagplot sono riportate le misure bivariate di tendenza centrale (media, mediana, ecc.) nonché una regione più scura ed una più chiara ed esterna intorno ad esse. Nel caso di tendenza centrale rappresentata da una mediana, la regione scura potrebbe rappresentare i valori compresi nell'intervallo di valori più prossimi alla mediana (ad es., intervallo definito dal 25-imo e il 75-imo percentile), mentre il "recinto" che delimita la regione più chiara potrebbe rappresentare l'area delimitata, ad esempio, dal 15-imo ed il 85-imo percentile. Le osservazioni fuori dal recinto saranno considerate outlier.

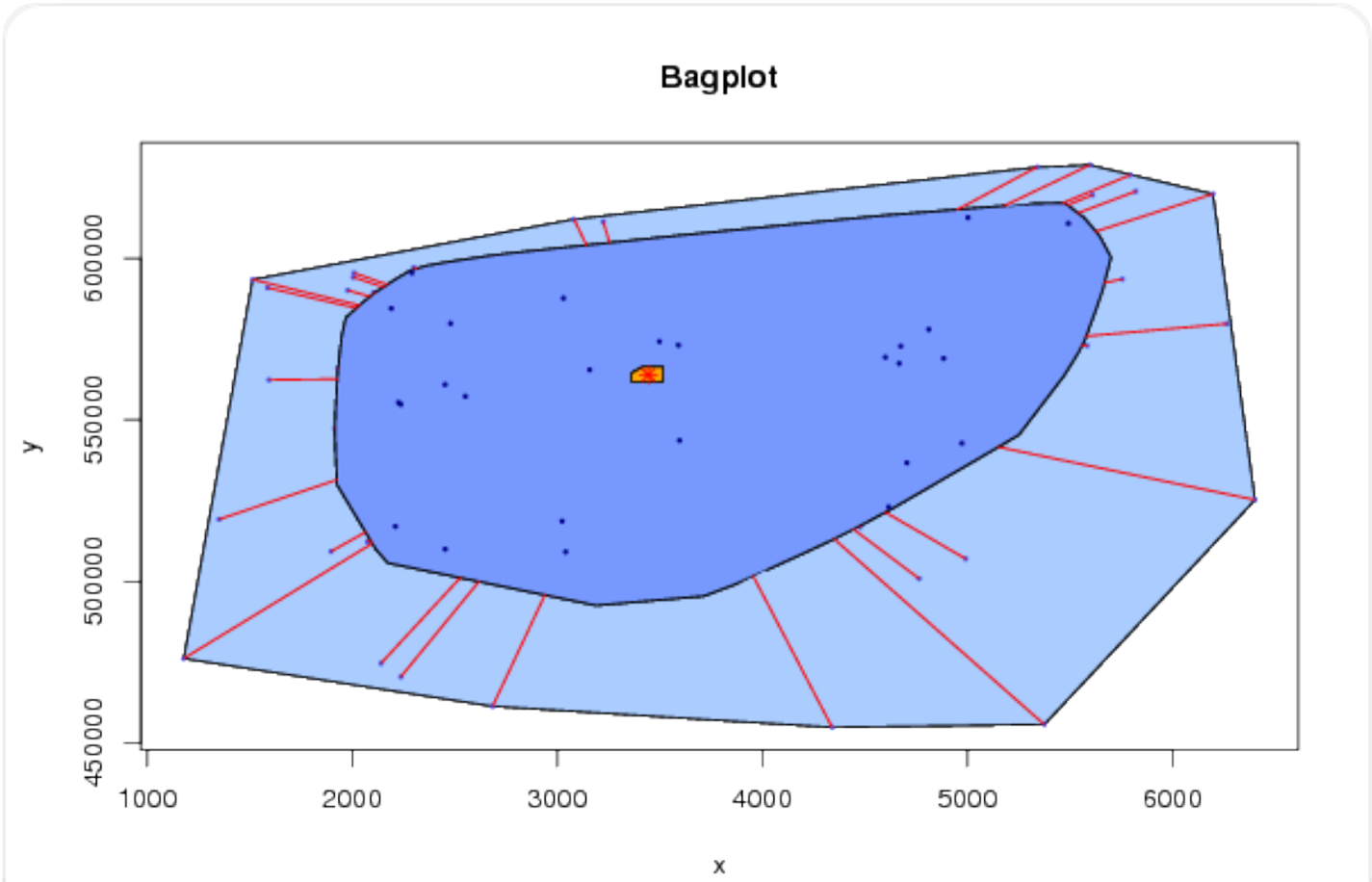


Fig. 7.2 Bagplot creato con Wessa

Tramite Wessa è possibile creare i bagplot gestendone ogni suo aspetto estetico e di contenuto (fig. 7.2; cliccare su *Compute*). Per la funzione particolare implementata in Wessa si ricorre all'utilizzo del pacchetto di R: *rpart*.

Capitolo 8

Confronto temporale

Grafico a linee

Sparkline

8

Confronto temporale tra variabili quantitative

Il **grafico a linee** (Harary & Norman, 1960) per più categorie è bene interpretabile soprattutto quando sono disponibili pochi step temporali (intervalli), solitamente riportati sull'asse orizzontale dal meno recente al più recente. In tali occasioni è possibile tracciare una serie di linee, passanti attraverso i diversi punti-dato di ogni categoria. In questo modo sarà possibile confrontare agevolmente l'evoluzione dei dati di tali categorie nel tempo.

Tra le visualizzazioni disponibili in Datawrapper vi sono le line chart. Per la loro realizzazione è possibile ricorrere all'utilizzo della caratteristica interfaccia utente a 4 step. Questo tipo di visualizzazione rappresenta lo strumento ideale per la creazione di grafici a linee (fig. 8.1).

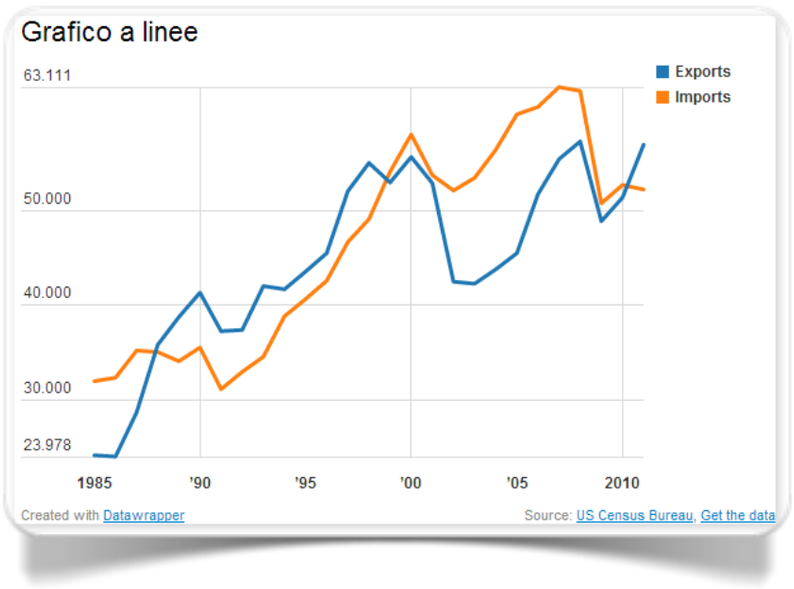


Fig. 8.1 Grafico a linee creato Datawrapper

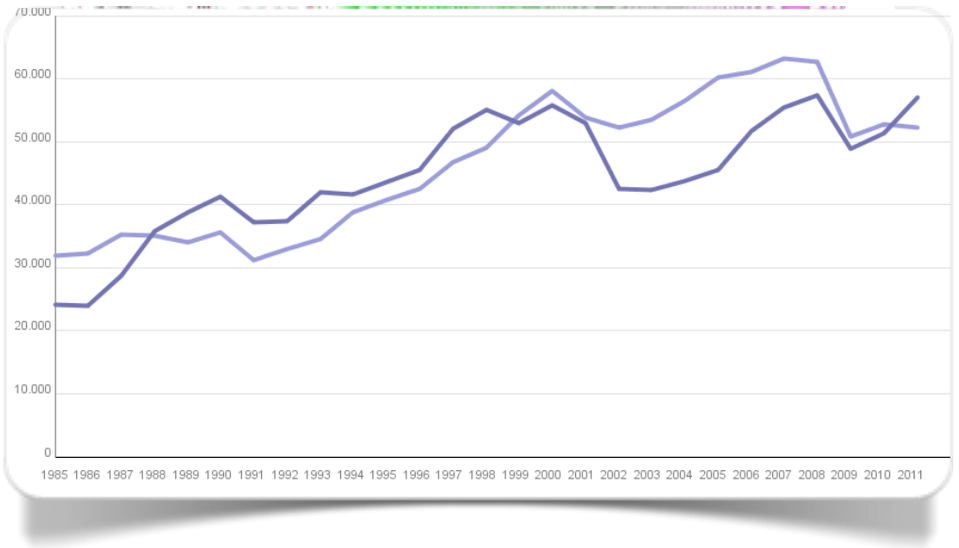


Fig. 8.2 Grafico a linee creato Many Eyes

Many Eyes consente di costruire grafici a linee ricorrendo alla visualizzazione Line Graph. Tra le opzioni disponibili vi è anche la possibilità d'impostare Relative: Set Start = 100, molto utile soprattutto quando si desidera rendere confrontabili diverse misurazioni indicizzando pari a 100 il valore corrispondente al primo step di confronto.

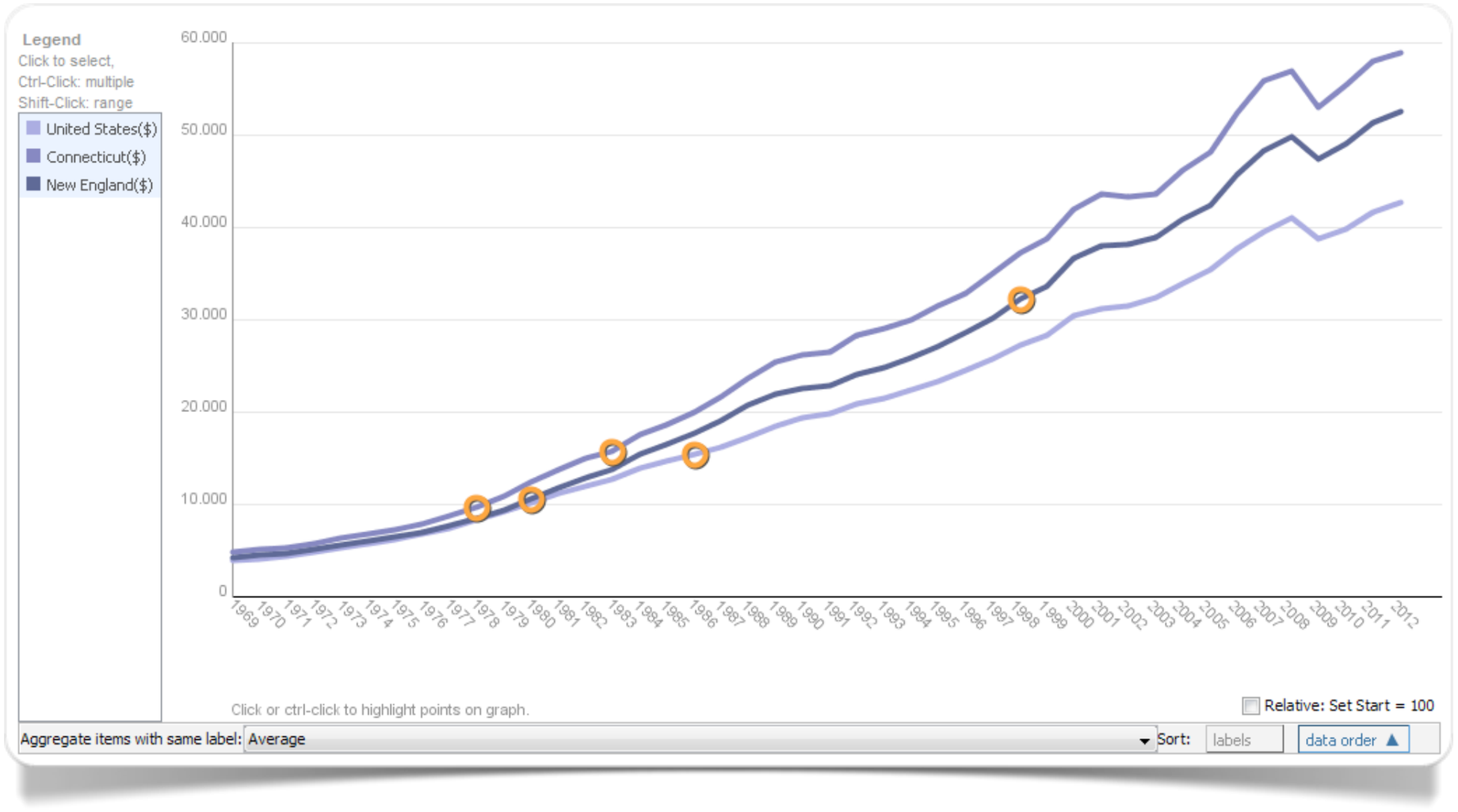


Fig. 8.3 Reddito pro capite in USA dal 1969 al 2012 (elab. Many Eyes)

Nella figura 8.3 vediamo una rappresentazione grafica della evoluzione del reddito pro capite negli USA (fonte DECD-CT).

Confronto temporale: serie storica

Quando si desidera riportare su grafico i dati di una serie storica, lo strumento certamente più adeguato è rappresentato da un **grafico a linee** (Harary & Norman, 1960). Le serie storiche sono solitamente caratterizzate da un numero elevato di step temporali in corrispondenza dei quali è disponibile un corrispettivo valore numerico (la maggior parte delle volte di natura continua). Il grafico a linee è l'unico tipo di visualizzazione in grado di evidenziare tutte le componenti principali di una serie storica (tendenza, ciclicità, stagionalità, ecc.).

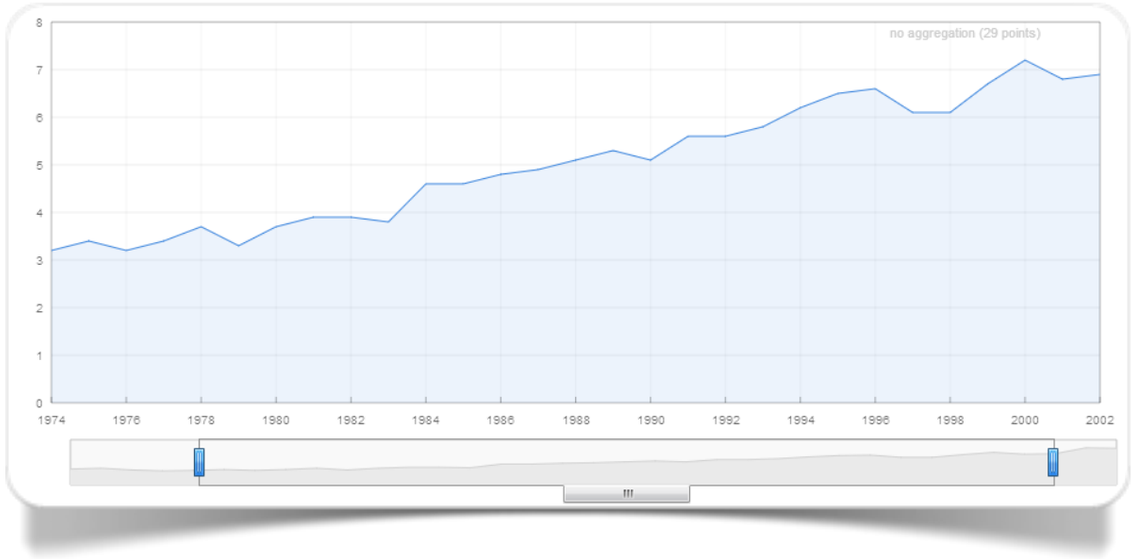


Fig. 8.4 Grafico a linee creato con ManyEyes

Many Eyes consente di costruire grafici a linee ricorrendo all'opzione di visualizzazione *View in Contest*. Soluzione particolarmente adatta quando sono presenti molti step temporali, in quanto presenta una funzionalità di zoom disponibile sulla parte del grafico: tramite di essa è possibile infatti selezionare sotto-intervalli temporali per poter visualizzare solo singole porzioni della serie storica (fig. 8.4).

Confronto temporale: serie storica ad alta intensità

Un grafico **sparkline** (Tufte, 2004) è generalmente contraddistinto da due principali caratteristiche: piccole dimensioni ed alta densità dei dati. Lo sparkline rappresenta trend e variazioni associate ad una particolare misurazione (temperatura, andamenti finanziari) nel modo più semplice possibile. In generale lo strumento di rappresentazione usato per riprodurre uno sparkline può essere un grafico a linee, uno scatterplot oppure un grafico a barre.

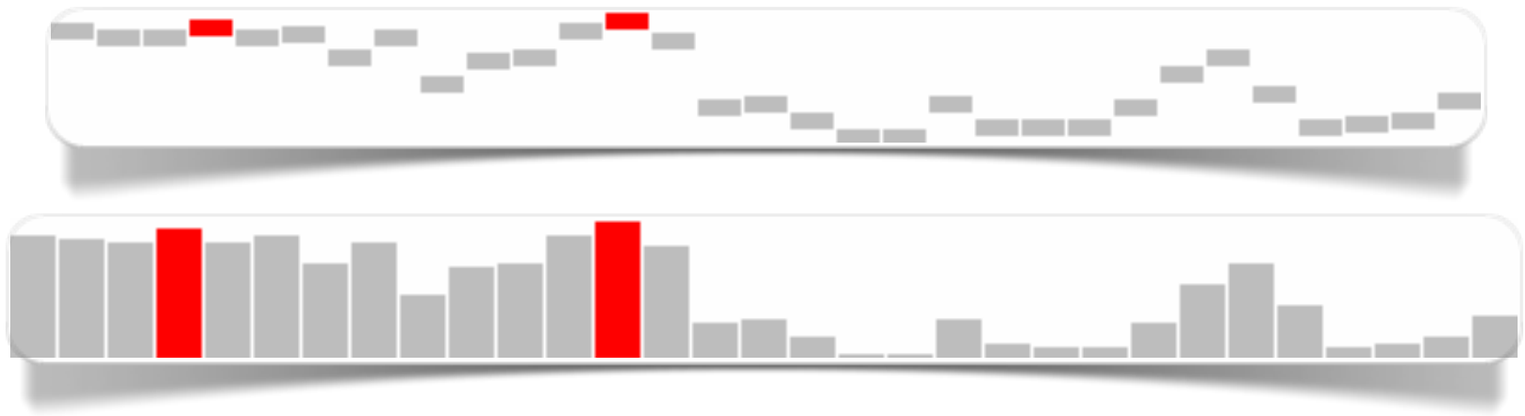


Fig. 8.5 Sparkline creati con Sparkline bitworking

Ideato da Joe Gregorio, **Sparklines bitworking** è uno strumento su web che consente di utilizzare Google Chart API per costruire facilmente sparkline a linee o a barre, controllandone tutti gli aspetti grafici (fig. 8.5).

Capitolo 9

Concentrazione

Curva di Lorenz

9

Concentrazione di una variabile quantitativa

La **curva di Lorenz** (Lorenz, 1905) è il principale strumento di rappresentazione degli indici di concentrazione. La curva è rappresentata in un piano sulla cui ascissa sono riportate le frequenze cumulate relative mentre sull'ordinata sono riportate le quantità cumulate relative. L'area compresa tra la curva e la retta di equidistribuzione (la retta a 45°) è detta area di concentrazione e può essere utilizzata come base per la definizione di appositi rapporti di concentrazione. Maggiore infatti è la concentrazione osservata, maggiore sarà tale area.

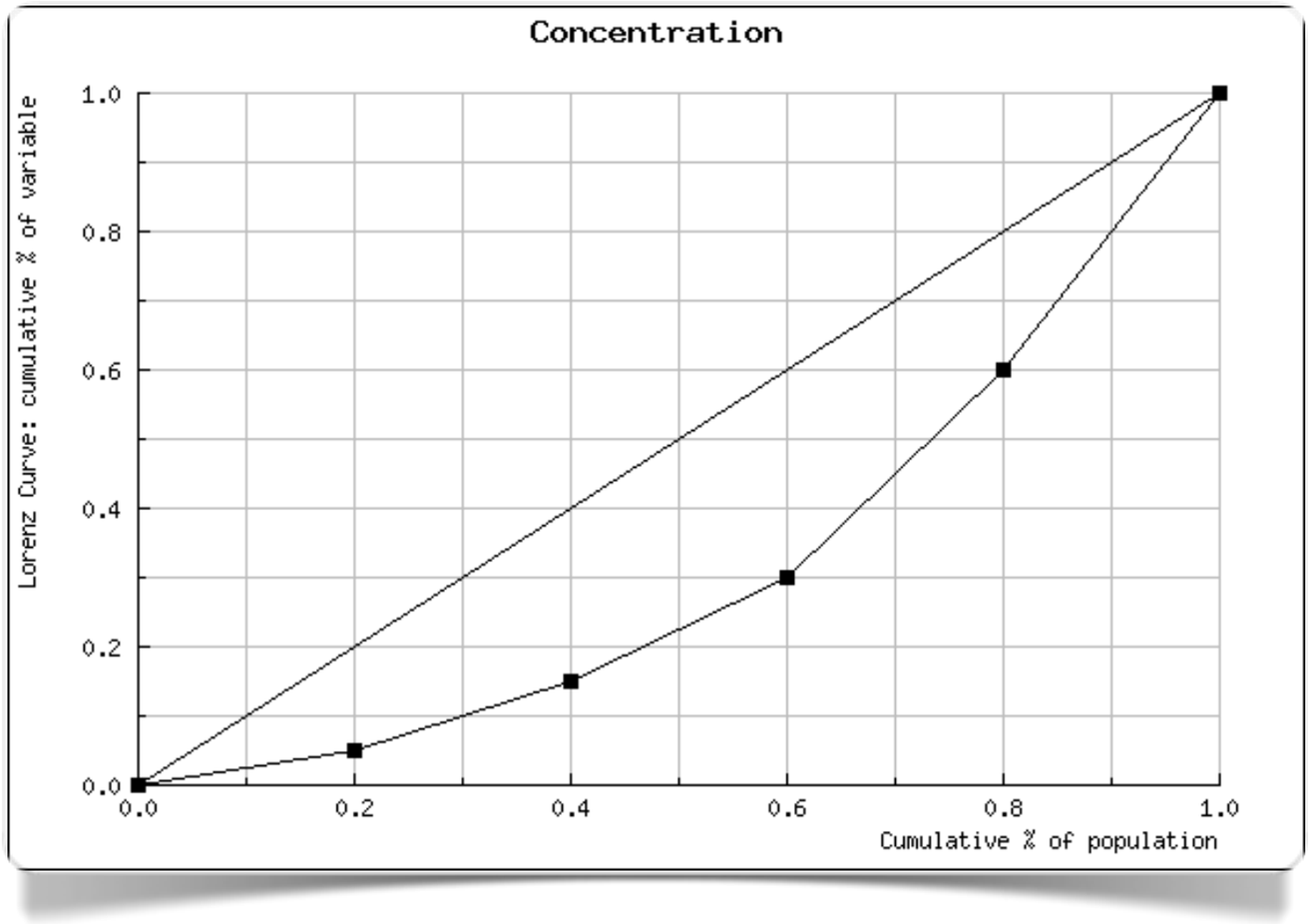
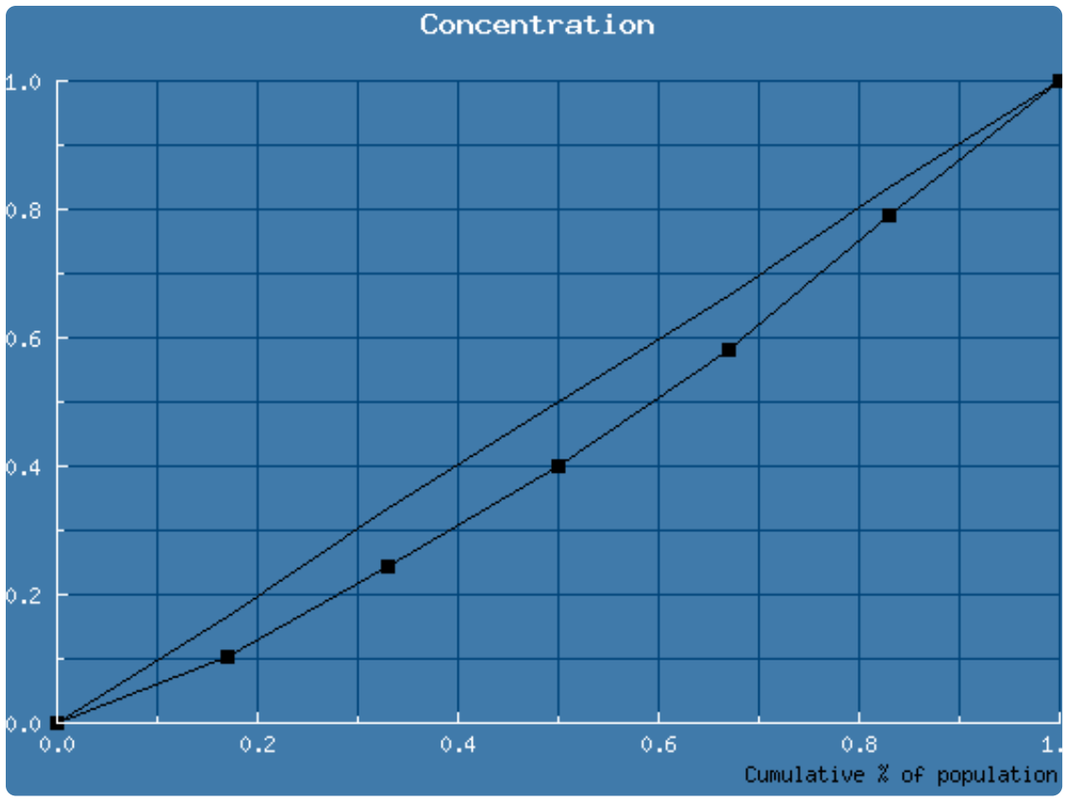


Fig. 9.1 Curva di Lorenz creata con Wessa

Questa curva di Lorenz o **grafico di concentrazione** (fig. 9.1) è stato realizzato con Wessa con i dati pre-inserirti nel campo *Data*.

Nella gallery 9.1 possiamo osservare la curva di Lorenz relativamente ai dati della figura a scorrimento 9.2 in cui abbiamo il PIL 2012 pro capite in dollari USA per alcuni paesi europei e per alcuni paesi africani (fonte: Internation Monetary Fund - Wikipedia).



Gallery 9.1 PIL pro capite 2012: confronto tra alcuni paesi europei e paesi africani

PIL pro capite in dollari per Germania, Francia, Regno Unito, Italia, Spagna, Grecia

Country	GDP 2012 per capita (US$)
Germany	44,513
France	44,141
United Kingdom	38,589
Italy	33,115
Spain	29,289
Greece	22,055
South Africa	7,507
Algeria	5,694
Tunisia	4,232
Egypt	3,113

Fig. 9.2 Dati sul PIL pro capite relativi ai grafici della Gallery 9.1 - Confronto tra la curva di concentrazione di paesi europei e paesi africani

Capitolo 10

Classificazione

Curva ROC

Dendrogramma

10

Classificazione di una variabile quantitativa

La **curva ROC** è uno strumento molto utilizzato in statistica biomedica. Nella sostanza si tratta della rappresentazione grafica di un classificatore binario, i cui due assi rappresentano generalmente la sensibilità ed il valore (1 – specificità) di un particolare test. La struttura dei dati richiede normalmente una **variabile** numerica di cui è identificato un valore soglia ed una seconda variabile a due categorie (ad es., positivo o negativo). La curva ROC consente di analizzare la performance del test lungo tutto l'intervallo di variazione dei valori della variabile numerica. Un'area sotto la curva (AUC) pari a 1 indica un test perfetto, mentre un'area pari a 0,5 (curva ROC equivalente a retta a 45°) indica un test che ha probabilità pari a 0,5 di classificare positivo.

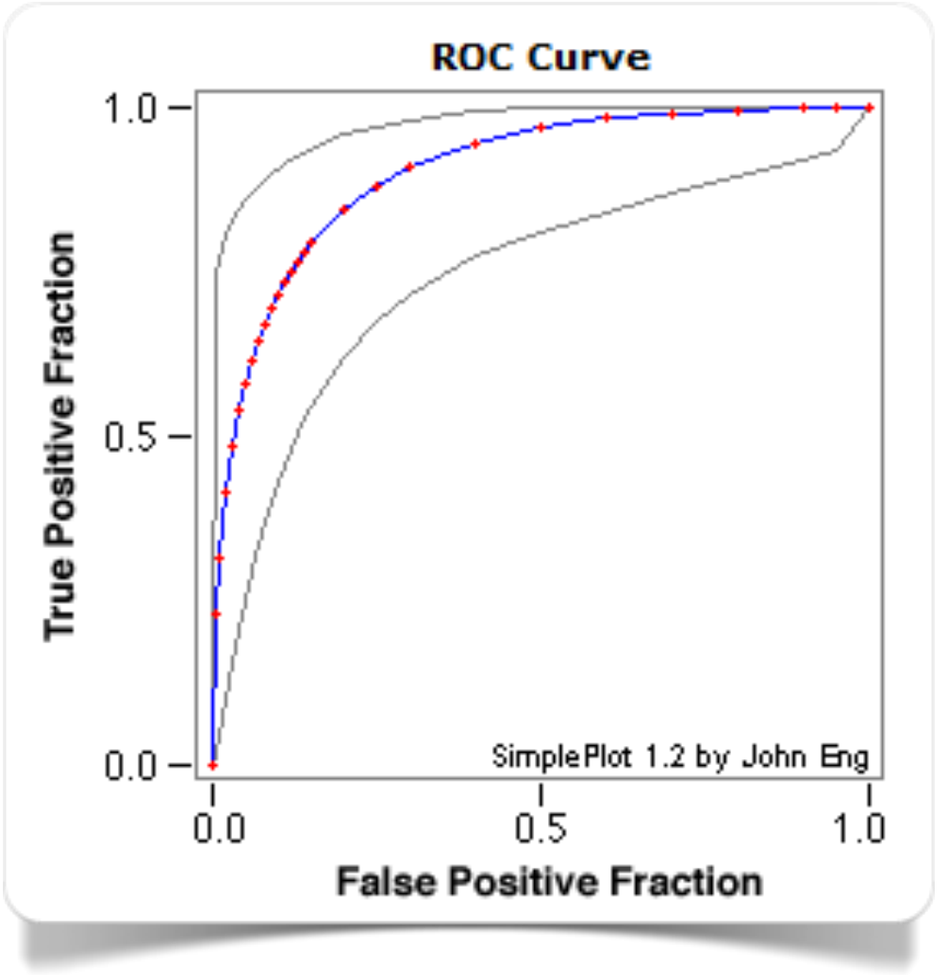


Fig. 10.1 Curva ROC creata con JROCFIT

JROCFIT è un software web messo a disposizione dalla Johns Hopkins University, Baltimore, Maryland, USA per consentire ai suoi studenti e non solo di produrre curve ROC. Nel portale viene spiegato che formato devono avere i dati così come vengono riportate le istruzioni su come esportare i risultati.

Su web è possibile ottenere un'analisi della Curva ROC anche attraverso Wessa , ed in particolare attraverso l'impiego della sua interfaccia basata sul software open source R (vedi http://www.wessa.net/rwasp_logisticregression.wasp).

Classificazione: analisi dei gruppi (raggruppamento gerarchico)

Il **dendrogramma** è il grafico utilizzato per rappresentare i risultati di un'analisi dei gruppi (*cluster analysis*) secondo la tecnica del raggruppamento gerarchico. Ogni gruppo è definito da minimo un membro (gruppo composto da un'unica osservazione) ad un massimo che equivale al numero totale di osservazioni (un unico gruppo contenente tutte le osservazioni). La distanza tra un estremo e l'altro del grafico definisce il grado di omogeneità dei membri appartenenti al medesimo gruppo. Quanto più prossima all'estremo di partenza (passo 0) è l'unione tra più osservazioni, tanto maggiore sarà il grado di omogeneità tra le osservazioni in termini di caratteristiche appartenenti al gruppo formatosi in seguito a tale unione.

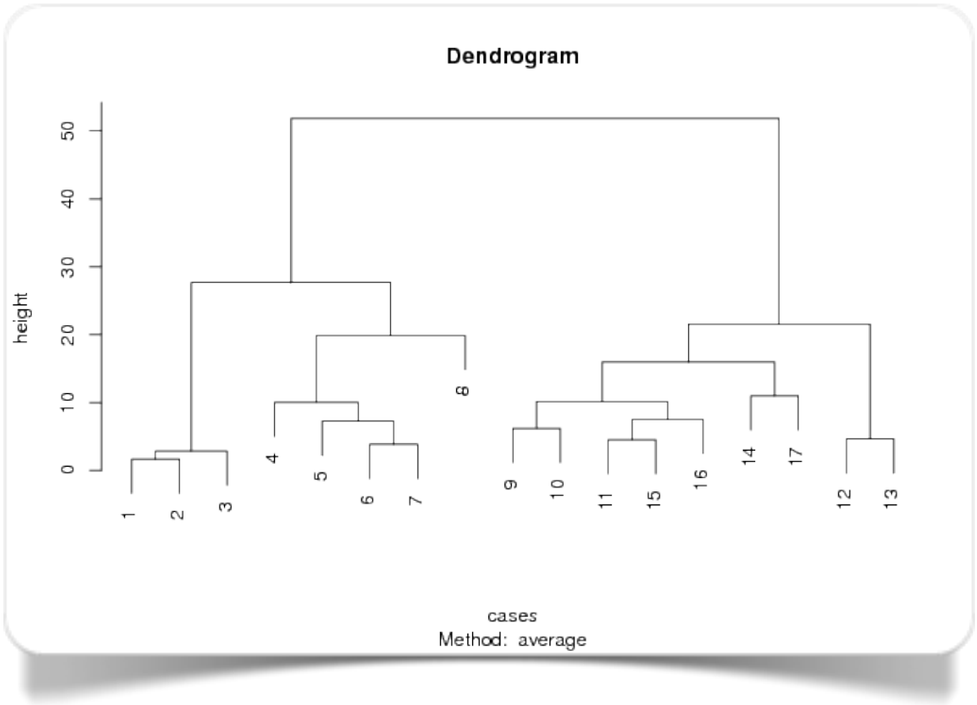


Fig. 10.2 Dendrogramma creato con Wessa

Con Wessa è possibile creare dendrogrammi di ogni livello di complessità (fig. 10.2; cliccare su *Compute*). Trattandosi di un grafico tradizionalmente realizzato nel contesto dell'analisi dei gruppi, per la realizzazione di questo grafico, si ricorre all'utilizzo del pacchetto di R: *cluster.*

Capitolo 11

Bibliografia e risorse in web

11

Bibliografia

Aitchison J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman & Hall, reprinted in 2003 with additional material by Caldwell, NJ: The Blackburn Press.

Chambers J. M., Cleveland W., Kleiner B., Tukey P. (1983) *Graphical Methods for Data Analysis.* Wadsworth International Group.

d'Ocagne M. (1885) *Coordonnées parallèles et axiales : Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles*. Paris: Gauthier-Villars.

Few S. (2006) *Information Dashboard Design: The Effective Visual Communication of Data*. Sebastopol, CA: O'Reilly Media.

Harary F., Norman R. Z. (1960) Some properties of line digraphs. *Rendiconti del Circolo Matematico di Palermo,* 9 (2): 161–169.

Lorenz M. O. (1905) Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, Vol. 9, No. 70: 209–219.

Pearson K. (1895) Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 186: 343–326.

Pearson K. (1904) On the Theory of Contingency and Its Relation to Association and Normal Correlation, in *Research Memoirs Biometric Series I*, Drapers' Company.

Playfair W. (1786) *The Commercial and Political Atlas: Representing, by Means of Stained Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure and Debts of England during the Whole of the Eighteenth Century*. London: Debrett; Robinson; and Sewell.

Playfair W. (1801) *Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe*. London: Wallis.

Rousseeuw P. J., Ruts I., Tukey, J. W. (1999) The Bagplot: A Bivariate Boxplot. T*he American Statistician.* 53 (4): 382–387.

Shneiderman B., Plaisant C. (2009) Treemaps for space-constrained visualization of hierarchies. Dec. 26th, 1998, last updated November, 2013; *Retrieved,* June 11, 2014.

Sneath P. H. A. (1957) The application of computers to taxonomy. *Journal of General Microbiology*, 17 (1): 201–226.

Tufte E. (2004) Sparkline theory and practice. Edward Tufte forum, May 27; *Retrieved,* June 11, 2014.

Tukey J. W. (1977) *Exploratory Data Analysis.* Boston: Pearson, Addison-Wesley

Venn J. (July 1880). On the Diagrammatic and Mechanical Representation of Propositions and Reasonings. *Philosophical Magazine and Journal of Science*. 5 10 (59).

Risorse in web

Datawrapper (https://datawrapper.de/)

Google Search (https://www.google.com/)

JROCFIT (http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html)

Many Eyes (http://www-958.ibm.com/software/data/cognos/manyeyes/)

plotly (https://plot.ly/)

R - The R Project for Statistical Computing (http://www.r-project.org/)

Sparklines bitworking (http://sparklines.bitworking.info/)

Raw (http://raw.densitydesign.org/)

Wessa (http://www.wessa.net/)

WolframAlpha (http://www.wolframalpha.com/)

SAPIENZA
UNIVERSITÀ DI ROMA

Dipartimento di Scienze statistiche

Data Science Series

1. L.Giuliano, *Il valore delle parole. L'analisi automatica dei testi in Web 2.0*.
2. D. Schiavon, L. Giuliano, *Wizard grafico. Una guida alla visualizzazione dei dati numeric*i
3. L. Giuliano. *The Value of Words. Automatic Text Analysis Tools in Web 2.0* (in preparazione)

Duccio Schiavon, Luca Giuliano
Wizard Grafico. Una guida alla visualizzazione grafica dei dati numerici
Roma : Dipartimento di Scienze statistiche, [2014] 69 p.

ISBN 978-88-908757-1-7

Composizione

Per composizione s’intende l’insieme di dati quantitativi che rappresentano ognuno una parte del tutto e che descrivono esclusivamente una parte relativa d’informazione (Aitchison,1986). Nella statistica, l’utilizzo di questo tipo di dati è frequente quando ogni punto-dato rappresenta una “frazione” di un insieme non negativo di numeri la cui somma è 1. In genere, ogni punto-dato suggerisce la proporzione (o “percentuale”) di unità statistiche che corrispondono a una specifica categoria all’interno dell’insieme totale di categorie presenti nell’insieme di dati.

Termini del glossario correlati

Trascina termini correlati qui

Indice

Trova termine

Curva interpolante

Per curva interpolante, s’intende la funzione derivante dal processo di *curve fitting*. Il *curve fitting* consiste nella costruzione di una curva o di una funzione matematica caratterizzata dalla migliore corrispondenza con una serie di punti.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Data journalism

Il *data journalism* (o *data-driven journalism*) si può considerare un particolare metodo di giornalismo basato sull'analisi di grandi insiemi di dati. Nella maggior parte dei casi si tratta di *open data* liberamente disponibili su web, e la loro elaborazione richiede l'impiego di strumenti *open source*.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Diagramma di flusso

Un diagramma di flusso è un tipo di diagramma che rappresenta un algoritmo in forma di flusso: ogni passaggio è contraddistinto da particolari elementi grafici (cerchi, quadrati, box, ecc.), ed ognuno di essi è relazionato agli altri attraverso connessioni e frecce che ne determinano la direzione logica di lettura. Generalmente questa rappresentazione schematica ha la funzione d’illustrare la soluzione a un dato problema.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Differenza assoluta

Il termine differenza assoluta di due numeri reali x e y è data dalla dalla formula |x-y|, e rappresenta la distanza di una retta reale tra i punti corrispondenti a x e y.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Differenza relativa

Le differenze relative vengono solitamente utilizzate per confrontare quantità considerate in termini di porzioni di “dimensioni”. Il confronto si basa su misure espresse in rapporti e non esprimibili sulla base di alcuna unità di misura. Se tali rapporti vengono moltiplicati per 100, tali rapporti possono essere considerati come valori percentuali. In questo caso le differenze relative possono considerarsi vere e proprie differenze percentuali.

Termini del glossario correlati

Trascina termini correlati qui

Distribuzione

Nella statistica, il concetto di distribuzione si riferisce principalmente alla forma di una distribuzione di probabilità e ha lo scopo di suggerire visivamente quale potrebbe essere il migliore modello statistico da adattare ai dati che formano la particolare forma distributiva. La distribuzione ha quindi la particolare funzione “grafica” di evidenziare quali potrebbero essere le particolari proprietà statistiche della popolazione a cui appartiene l’insieme di dati analizzati.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Infografica

L'infografica (*information graphic* o *infographic*) è una forma di rappresentazione dell’informazione in cui numeri e testo trovano una loro precisa collocazione in una forma visiva organizzata. Le tecniche utilizzate per ottenere questo tipo di rappresentazioni richiedono competenze grafiche ed informatiche, nonché non indifferenti qualità espositive.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Open data

Per *open data* (dati aperti) s’intende l’insieme di dati liberamente accessibili e privi di restrizioni all’utilizzo e alla riproduzione. Perché sia soddisfatta la caratteristica indispensabile di “apertura”, i dati non devono essere vincolati da brevetti o da altre forme di controllo che ne limitino la riproduzione. Le uniche restrizioni consentite si riferiscono all’obbligo eventuale di citazione delle fonti o alle modalità di modifica.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Tabella di contingenza

La tabella di contingenza (Pearson, 1904) è un particolare tipo di tabella in forma di matrice in cui è riportata la distribuzione di frequenza (multivariata) delle variabili coinvolte nell'analisi.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Variabile

In statistica, una variabile rappresenta una caratteristica che può assumere più di un insieme di valori a cui associare una misura numerica o una categoria classificatoria (ad es., reddito, età, peso, ecc. per le variabili numeriche oppure “professione”, “colore occhi”, “malattia”, ecc. per le variabili categoriali).

Le variabili numeriche si suddividono principalmente in due categorie:

- Variabili continue, che possono assumere un numero infinito di valori tra due valori distinti (es., pressione arteriosa, temperatura, ecc.)
- Variabili discrete, che assumono valori da un insieme finito o conteggiabile di valori (ad es., numero di figli, numero di gambe di un animale, ecc.)

Le variabili categoriali si suddividono in due categorie:

- Variabili nominali, in cui le modalità identificano specifiche categorie, cioè caratteristiche o qualità precise non ordinabili (es., sesso, razza, mezzo di trasporto, ecc.)
- Variabili ordinali, in cui le modalità identificano categorie che possono essere organizzate sulla base di una qualche relazione d’ordine o gerarchia (es., titolo di studio, grado di soddisfazione, ecc.)

Termini del glossario correlati

Trascina termini correlati qui

Variabilità

In statistica, la variabilità (anche detta dispersione statistica o variazione) misura il grado di dispersione di una variabile o distribuzione probabilistica. In particolare, un indice di variabilità (varianza, deviazione standard, intervallo interquantile, ecc.) serve per descrivere quanto i suoi valori sono distanti dalla rispettiva misura di tendenza centrale (media, mediana, rango medio, ecc..).

Termini del glossario correlati

Trascina termini correlati qui
