Datasaurus Dozen

Lab. 01 | Es. 04

OBIETTIVO

Obiettivo dell'esercizio

DURATA

Durata dell'esercizio

MODALITÀ DI SVOLGIMENTO

A casa o in classe, a gruppi o singolarmente

INFORMAZIONI

→ Alberto Cairo

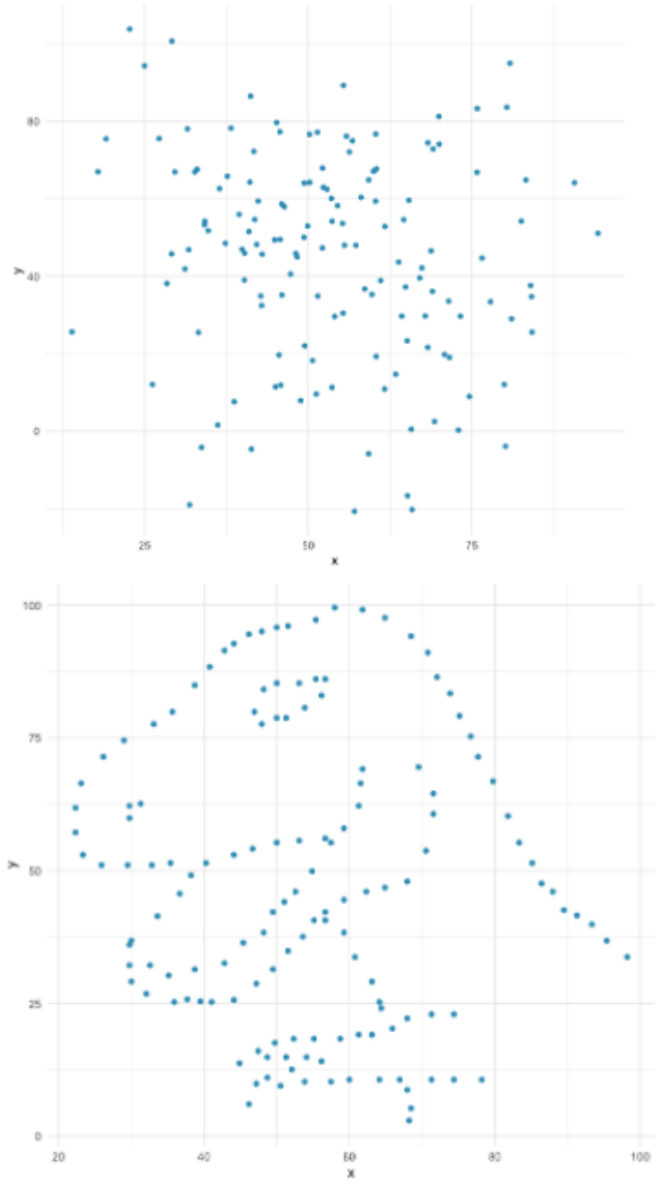
A questo punto entra in scena un nuovo protagonista: **Alberto Cairo**, uno dei più importanti esperti al mondo di Data Visualization, docente all'Università di Miami e autore di numerosi testi fondamentali sull'argomento.

Quando, nel 2016, Cairo vide DrawMyData ne rimase piacevolemente colpito e, oltre a commentare con un "Fantastic!" che fa bella mostra di sé sul sito di Grant, decise di utilizzare subito questo strumento per creare il suo più famoso dataset sintetico, il Datasaurus. Cairo pubblicò il suo **Datasaurus** sia su Twitter che sul suo blog per rinforzare il suo famoso concetto "Never trust summary statistics alone; always visualize your data!"

Come spiegato nell'Episodio 01, con questo concetto Cairo fa riferimento al fatto che, per quanto molto importanti e molto utili, le statistiche descrittive di un dataset non sono sufficienti, da sole, a spiegarne tutte le caratteristiche. Se prendiamo appunto il DataSaurus, le sue statistiche sono le seguenti:

N	media_x	sd_x	media_y	sd_y	P corr	beta
142	54.26	16.77	47.83	26.93	-0.0645	-0.1036

Se immaginiamo che questi descrittori provengano da un dataset generato da una distribuzione normale bivariata, l'aspetto che potremmo aspettarci potrebbe essere molto diverso dalla peculiare forma del DataSaurus:

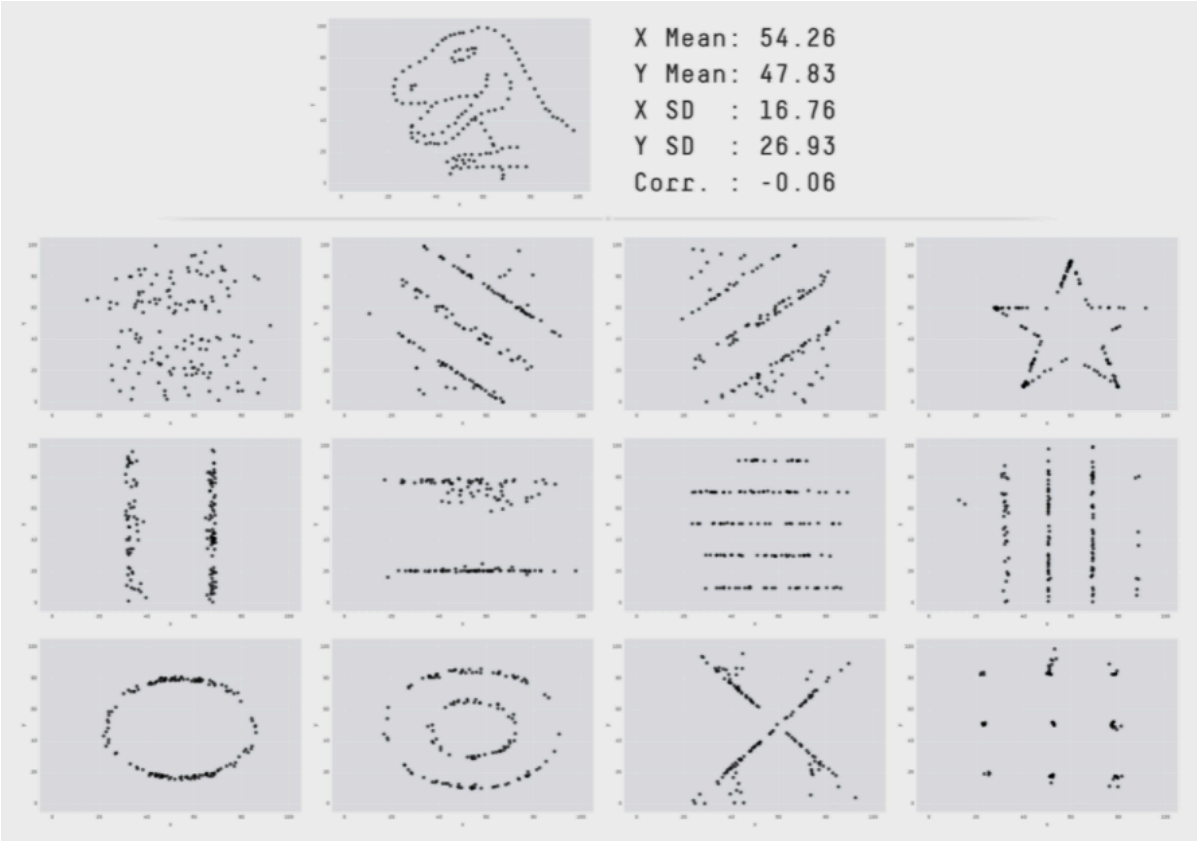


➜ Justin Matejka e George Fitzmaurice

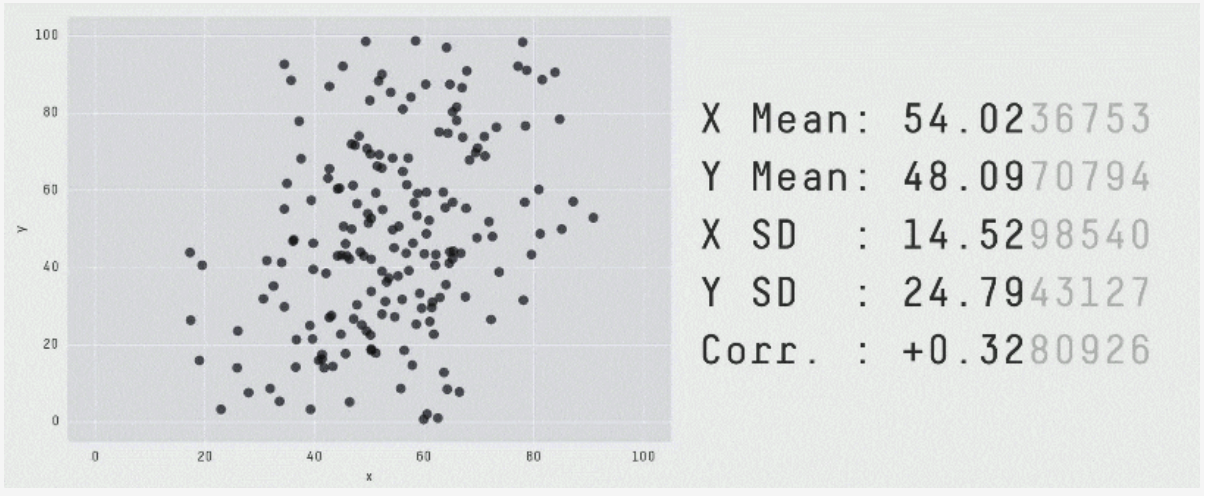
Arriviamo ora all'ultima parte del nostro viaggio: nel **2017** due ricercatori canadesi, **Justin Matejka** e **George Fitzmaurice** vinsero una 'Honorable Mention' presentando all'*ACM SIGCHI Conference on Human Factors in Computing Systems* una nuova suite di 12 dataset con le stesse statistiche descrittive del Datasaurus, generati tramite una strategia di ottimizzazione

derivata dalla fisica. Il relativo articolo, dal titolo "**Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing**", lo potete leggere qui e, se volete, potete anche esplorare il sito web che accompagna l'articolo.

I 12 dataset creati da Matejka e Fitzmaurice sono noti come **The Datasaurus Dozen** e hanno la seguente rappresentazione grafica:



In dettaglio, la costruzione dei 12 dataset è realizzata muovendo i punti del DataSaurus originale verso il tipo di forma desiderato, e ottimizzando le statistiche descrittive tramite un processo chiamato **Simulated Annealing**, che sfrutta le leggi della termodinamica, dove un parametro chiamato temperatura guida le variazioni delle coordinate dei punti a ogni iterazione. Qui di seguito, per esempio, vediamo come una nuvola di punti venga trasformata in un cerchio mantenendo invariate le statistiche:



La significatività di questi dataset rispetto all'originale quartetto di Anscombe risiede sia nel numero di dataset prodotti (12), che nella numerosità dei punti di ognuno dei dataset (ve ne sono diverse versioni, tutte con più di 100 punti).

I dataset inclusi nella Datasaurus Dozen possono essere scaricati qui, mentre ai due link di seguito si possono trovare i codici Python e R per provare a sperimentare con questi insiemi di dati se avete familiarità con questi due linguaggi

- Python
- R

ISTRUZIONI

➜ Step 1

Costruire un Foglio Google simile a quello fatto nel Lab.01 | Es.01 col quartetto di Anscombe e provare a modificare, eliminare, aggiungere punti e vedere l'effetto sulle relative statistiche e sull'aspetto grafico.

Consiglio

Per esempio, potreste provare a diminuire il numero di punti in modo omogeneo in tutti i dataset della DataSaurus Dozen e vedere quanto velocemente cambiano le statistiche descrittive.