Coding e Big Data

2023-2024

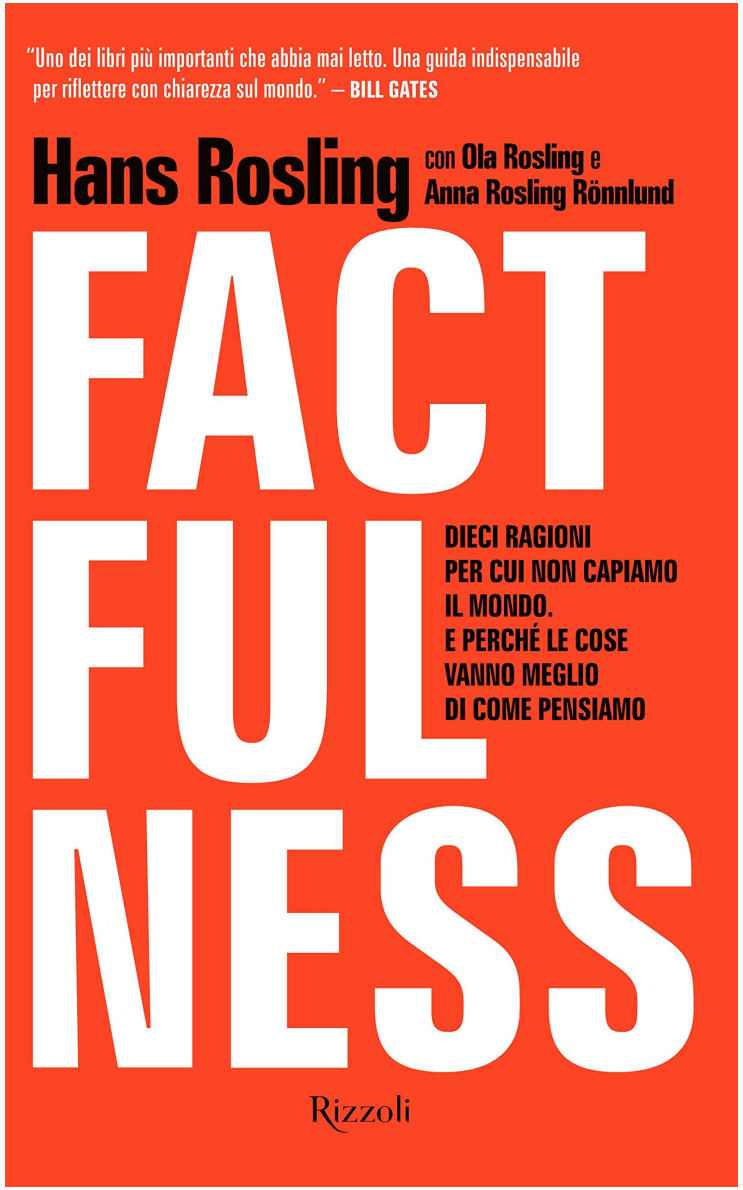
Vincenzo Nardelli

vincenzo.nardelli@unicatt.it

vincnardelli.com/cbd

“The world cannot be understood without numbers.
But the world cannot be understood with numbers alone.”

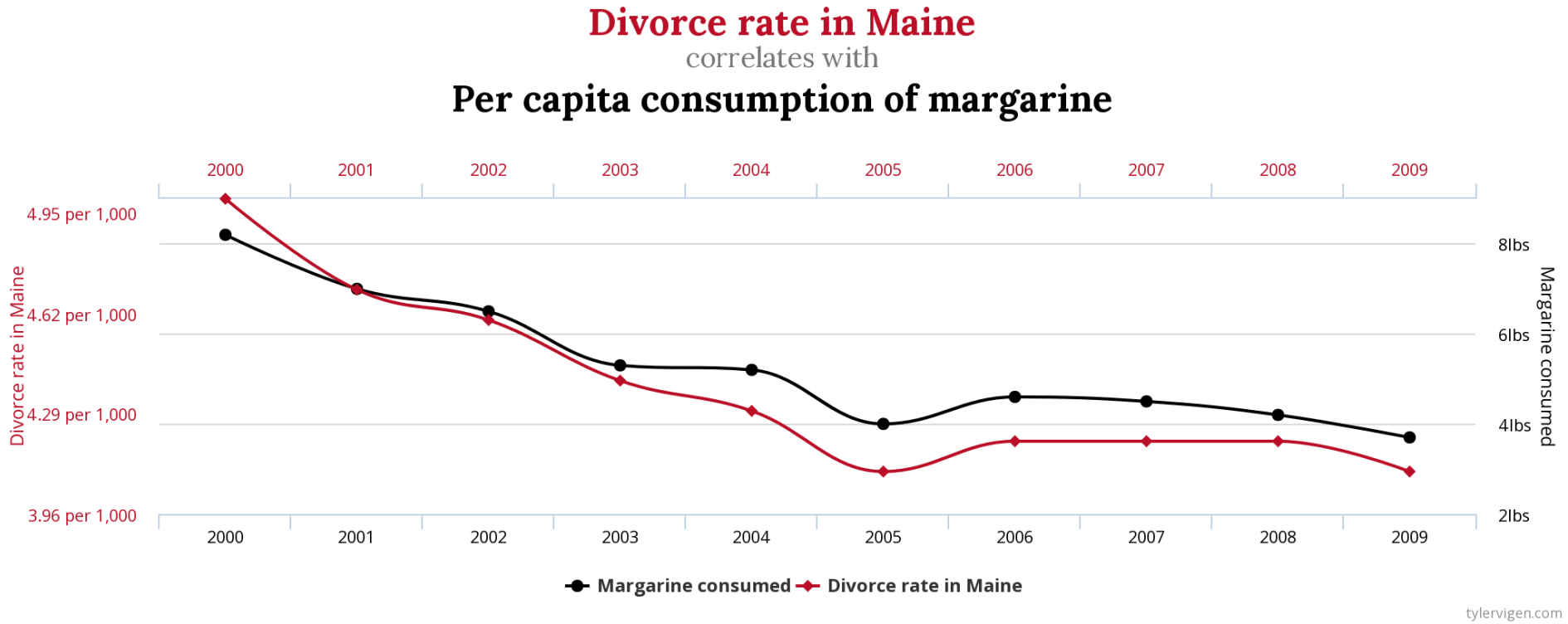
Hans Rosling



Dieci ragioni per cui non capiamo il mondo:

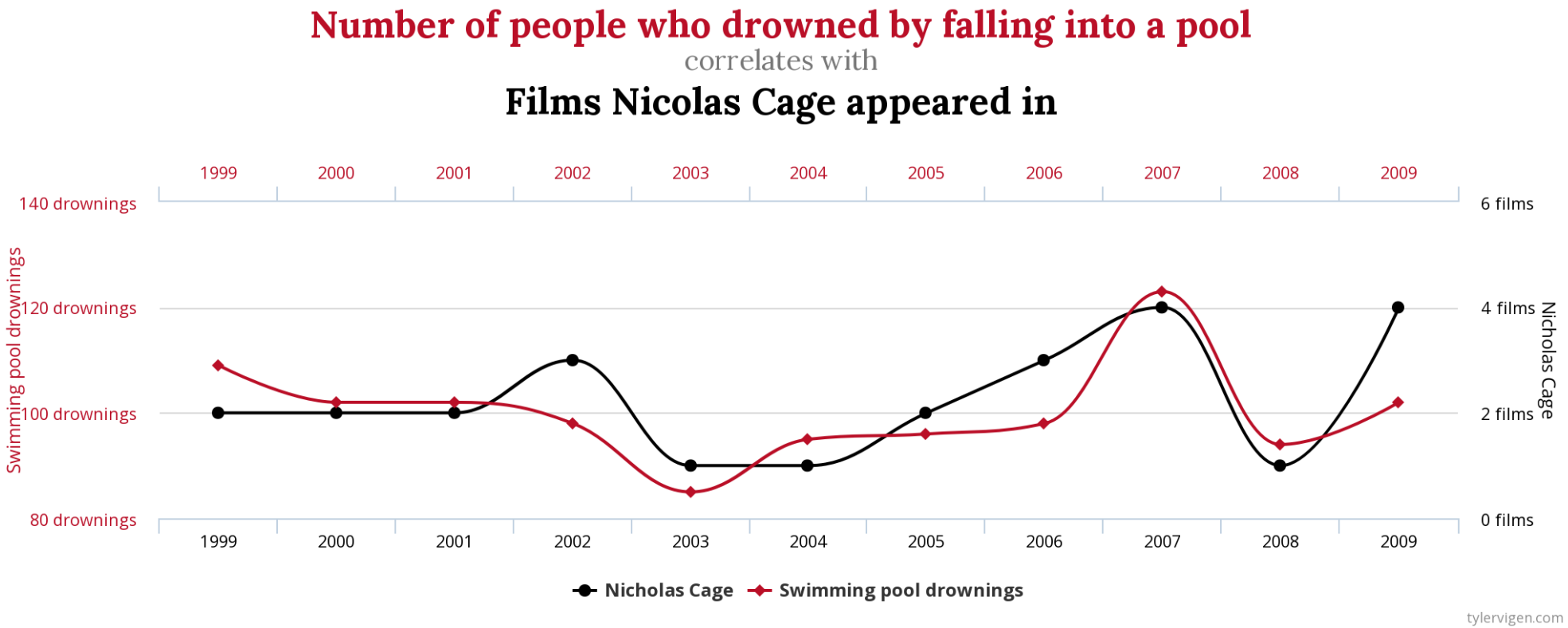
- Istinto del divario
- Istinto della negatività
- Istinto della linea retta
- Istinto della paura
- Istinto delle dimensioni
- Istinto della generalizzazione
- Istinto del destino
- Istinto della prospettiva singola
- Istinto dell'accusa
- Istinto dell'urgenza

Correlation is not causation



Altri esempi di
correlazioni spurie
https://www.tylervigen.com/spurious-correlations

Correlation is not causation



Altri esempi di
correlazioni spurie
https://www.tylervigen.com/spurious-correlations

Ricchezza in Italia

will_ita È la prima volta nella storia che i giovani sono più poveri dei propri genitori. Nel passato, le nuove generazioni avevano sempre più opportunità di migliorare le proprie condizioni socio-economiche rispetto a quelle precedenti. In Italia questo meccanismo si è bloccato. L'ascensore sociale è fermo, la mobilità intergenerazionale interrotta. In altre parole, è sempre più difficile migliorare la propria condizione sociale nel corso della vita.

Link

È corretto?
Chiediamo alle Intelligenze Artificiali

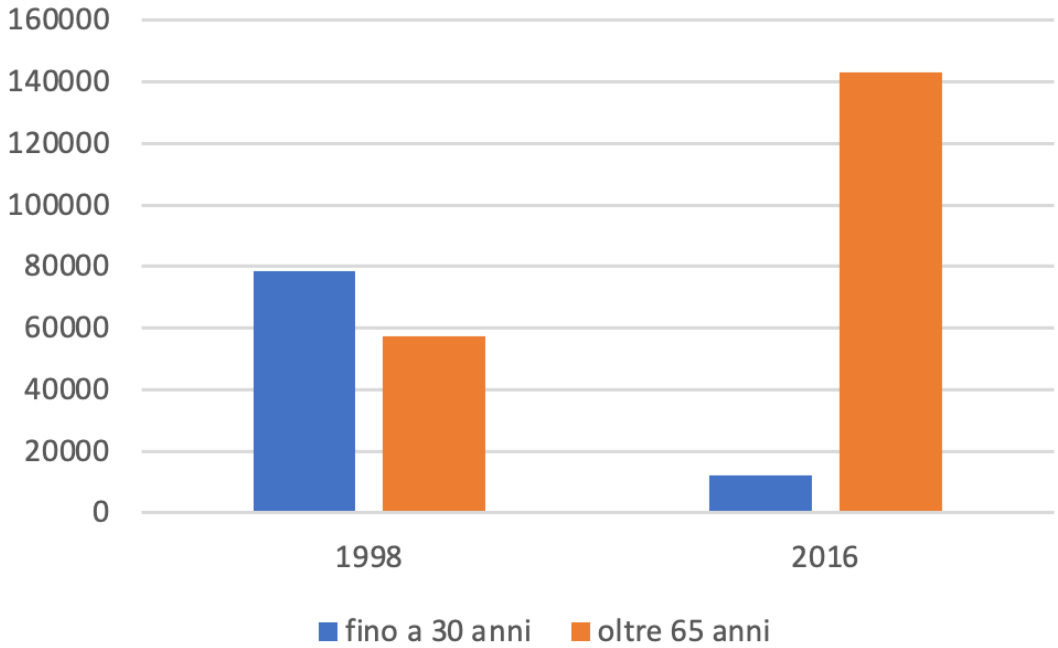
Link alla chat completa

Link alla chat completa

Fonte dati

- Indagine sui bilanci delle famiglie italiane nell'anno 2020 – Banca d'Italia https://www.bancaditalia.it/pubblicazioni/indagine-famiglie/bil-fam2020/index.html
  - TAVOLA S40

Analizziamo insieme i dati



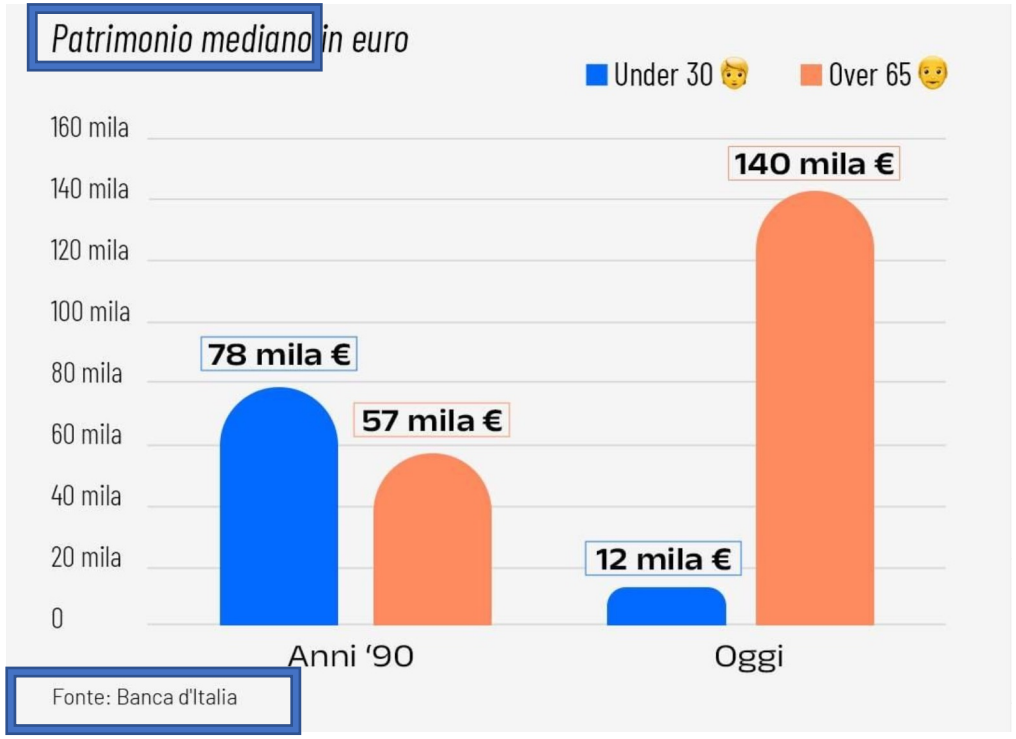
Processo di analisi dati

- Raccolta e selezione dei dati
- Pulizia dei dati
- Analisi esplorativa/descrittiva
- Modellistica
- Presentazione dei risultati

Processo di analisi dati

Raccolta e selezione dei dati

Attenzione ai dati ma anche ai **metadati**.



Risultati

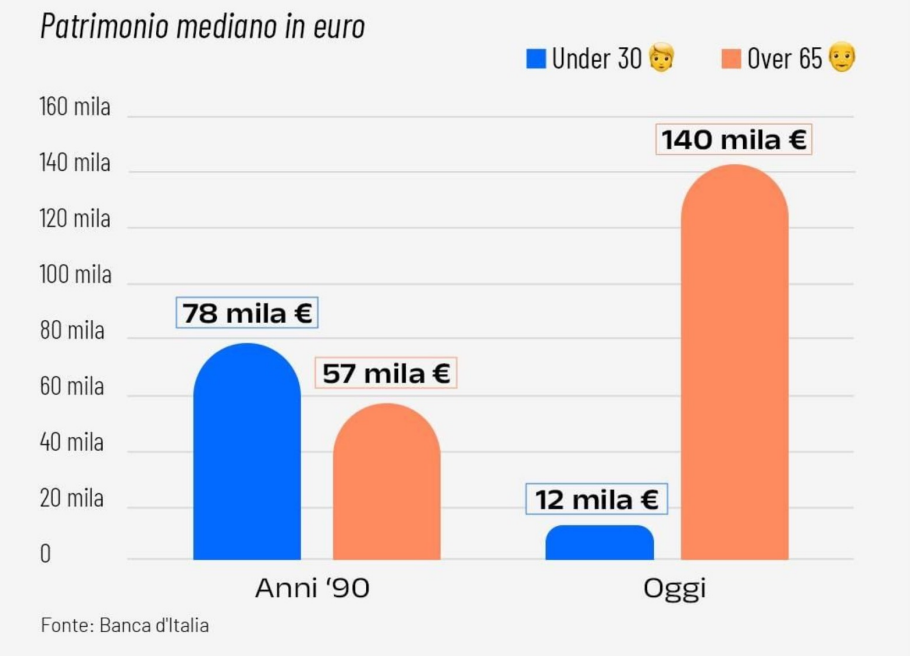
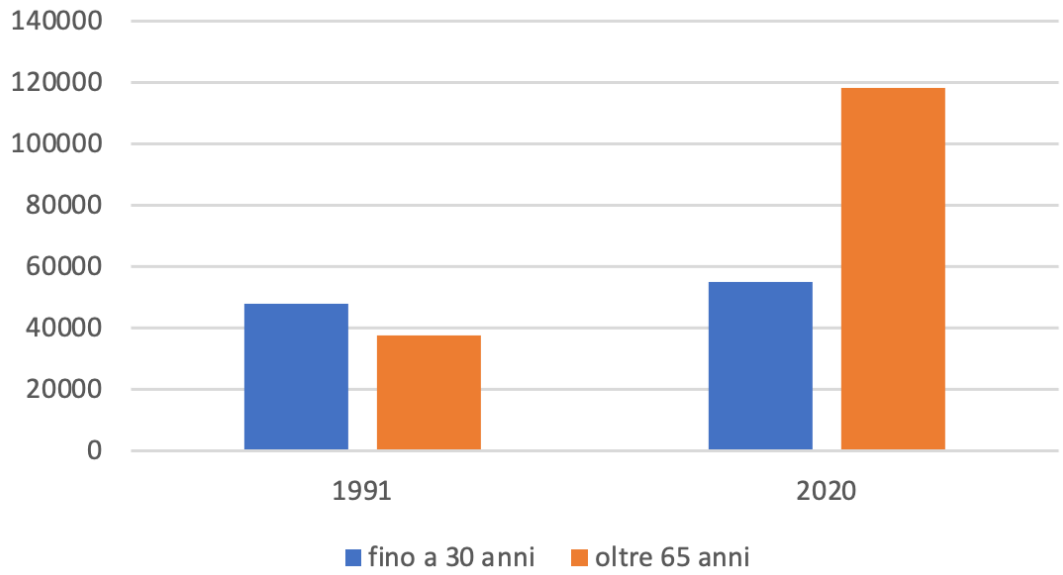


Grafico corretto



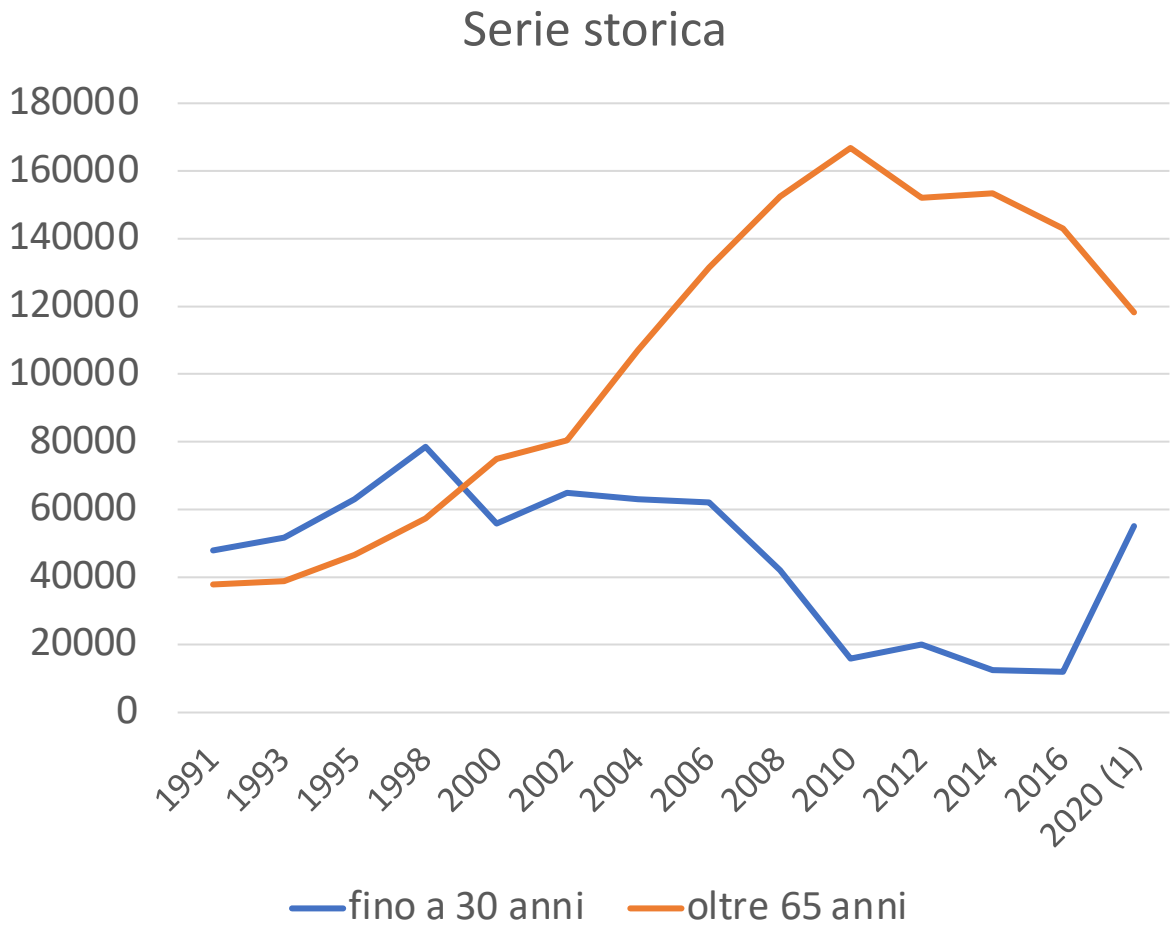
Risultati

will_ita È la prima volta nella storia che i giovani sono più poveri dei propri genitori. Nel passato, le nuove generazioni avevano sempre più opportunità di migliorare le proprie condizioni socio-economiche rispetto a quelle precedenti. In Italia questo meccanismo si è bloccato. L'ascensore sociale è fermo, la mobilità intergenerazionale interrotta. In altre parole, è sempre più difficile migliorare la propria condizione sociale nel corso della vita.

FALSO

Lo era fino al 2016.

Ora le cose stanno andando meglio!



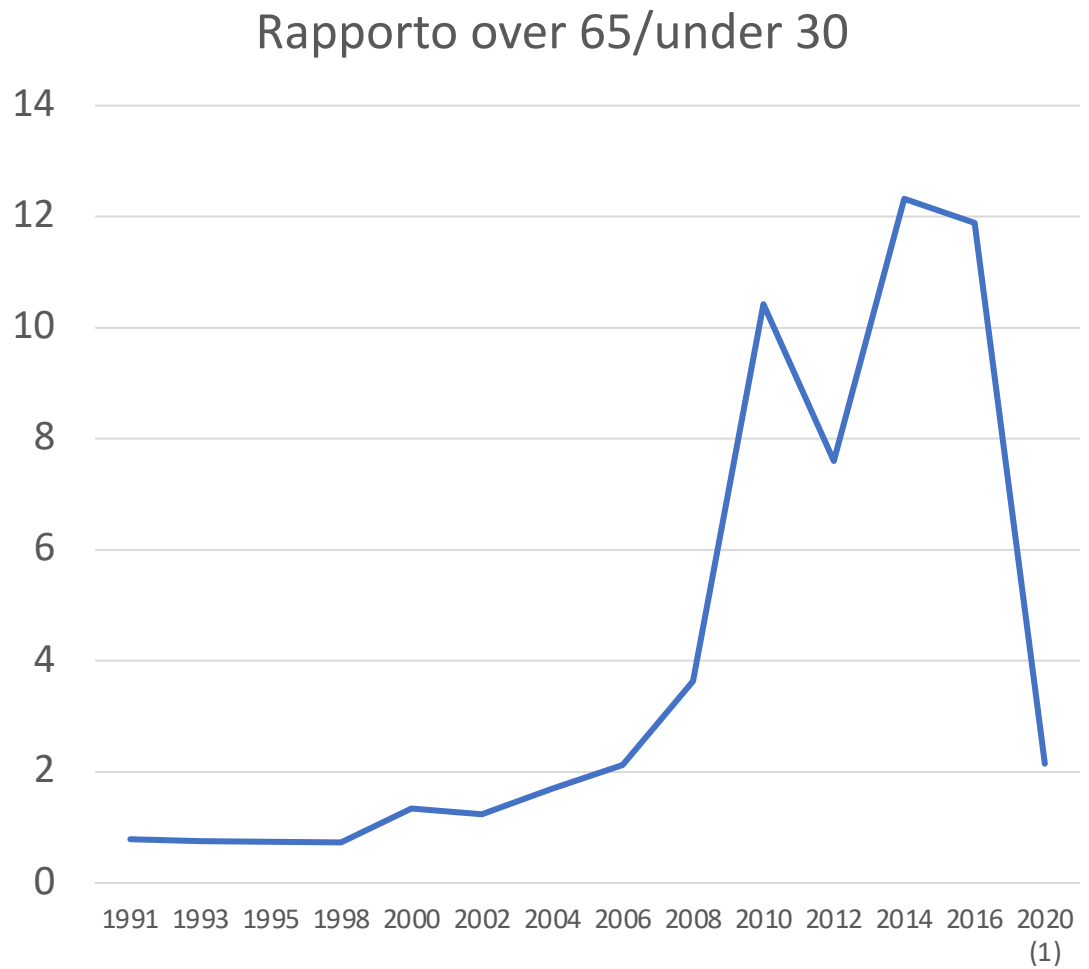
Risultati

In Italia gli over 65 sono 12 volte più ricchi degli under 30

FALSO

Lo era fino al 2016.

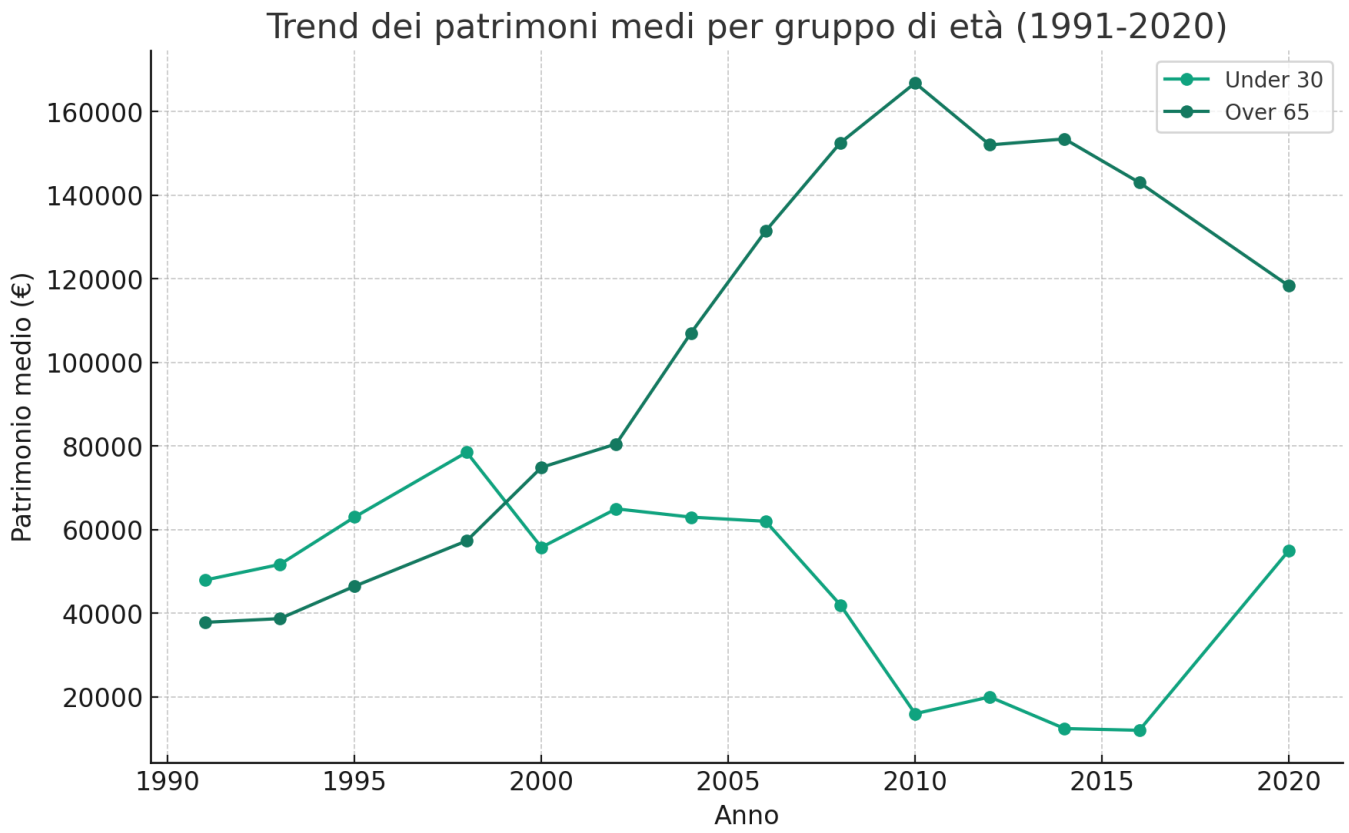
Ora le cose stanno andando meglio!



Ora facciamolo fare a Chat GPT

OpenAI

Link alla chat completa



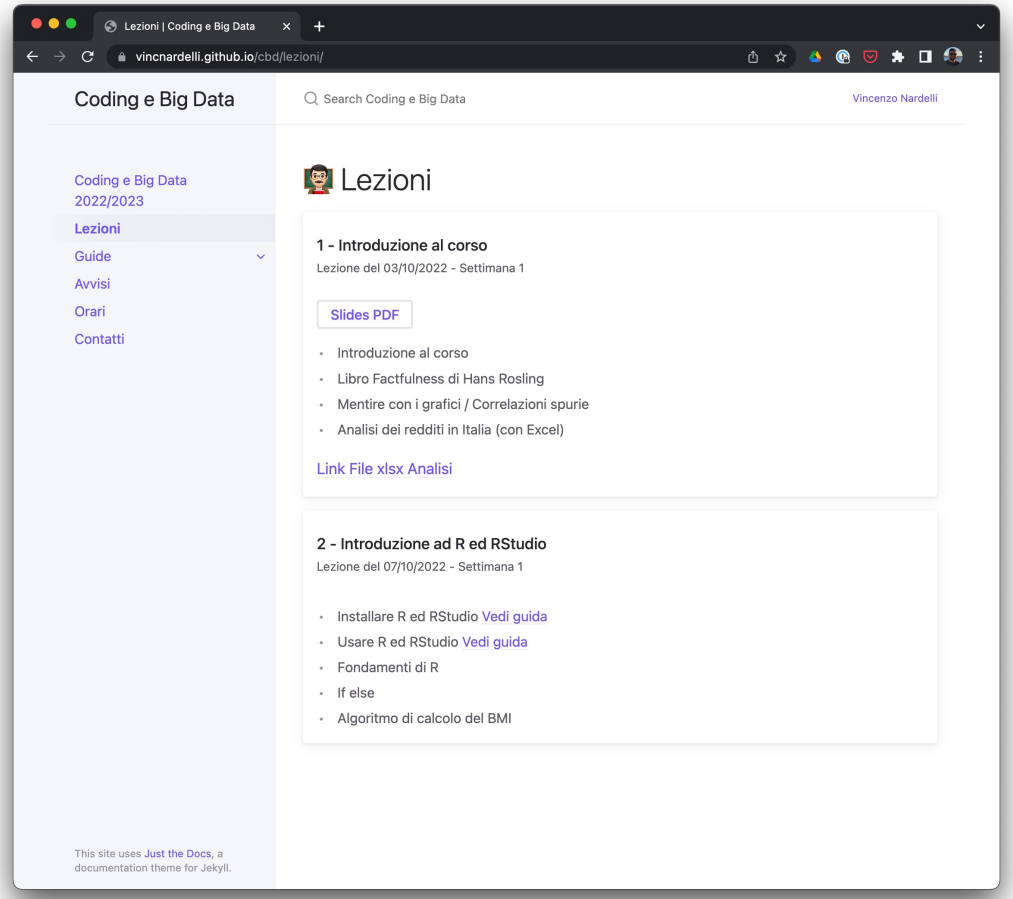
“Forming your worldview by relying on the media would be like forming your view about me by looking only at a picture of my foot.”

Hans Rosling

"Forming your worldview by ~~relying on the media~~ not relying on data analysis would be like forming your view about me by looking only at a picture of my foot."

Hans Rosling

Sito web del corso



https://vincnardelli.com/cbd/

Analisi di big data e programmazione

- Efficienza
  - Le interfacce punta e clicca non sono efficienti in termini di tempo
  - Automatizzare significa velocizzare le operazioni

- Riproducibilità
  - Crescente necessità di fornire dati, materiali ed analisi insieme ad i risultati
  - Assicura la possibilità di controllare i risultati e le procedure
  - Rende possibile effettuare analisi in produzione

Linguaggi di programmazione più usati

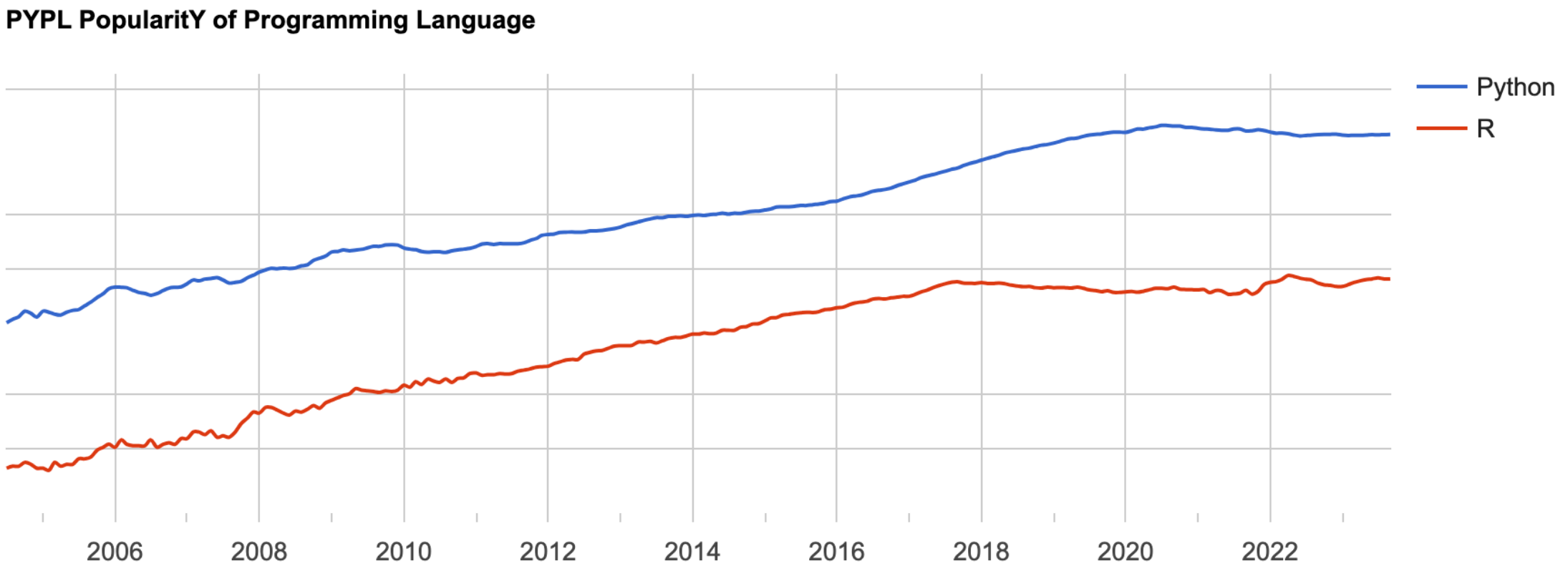
Worldwide, Sept 2023 :

Rank	Change	Language	Share	1-year trend
1		Python	27.99 %	+0.1 %
2		Java	15.9 %	-1.1 %
3		JavaScript	9.36 %	-0.1 %
4		C#	6.67 %	-0.4 %
5		C/C++	6.54 %	+0.3 %
6		PHP	4.91 %	-0.4 %
7		R	4.4 %	+0.2 %
8		TypeScript	3.04 %	+0.2 %
9	↑↑	Swift	2.64 %	+0.6 %
10		Objective-C	2.15 %	+0.1 %

Fonte

https://pypl.github.io/PYPL.html

Linguaggi di programmazione più usati



R & Python

- Open source
  - Gratuito e liberamente utilizzabile

- Strumenti avanzati
  - Pacchetti e librerie per ogni tipo di analisi

- Documentazione e comunità
  - Nessun supporto cliente a pagamento ma comunità!

R vs Python

- Data analytics, statistica
- Usato da statistici e dalla ricerca
- 12000 package on CRAN
- Semplice comunicazione (visualization, reporting and dashboard)

- Deployment and production
- Usato da programmatori e sviluppatori
- Integrazione con diversi sistemi operativi
- Algoritmi complessi e struttura ad oggetti

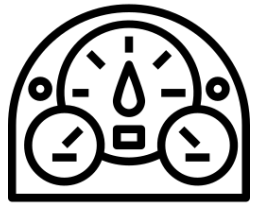
R e RStudio

Linguaggio

Motore

IDE

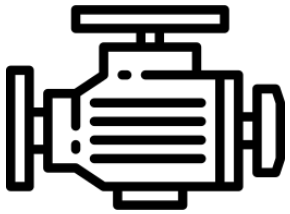
(integrated development environment)



Cruscotto

Python

Linguaggio



Motore

IDE

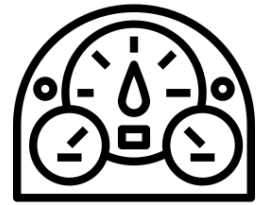
(integrated development environment)

jupyter

Visual Studio Code

PC

SPYDER



Cruscotto

Curva di apprendimento

EXCEL

La maggior parte delle persone probabilmente ha già appreso almeno alcune nozioni di base in Microsoft Excel. Questo è un vantaggio sostanziale dell'utilizzo di Excel: la curva di **apprendimento iniziale è piuttosto minima** e la maggior parte delle analisi può essere eseguita puntando e facendo clic sul pannello superiore. Una volta che un utente ha importato i propri dati nel programma, non è eccessivamente difficile creare grafici e diagrammi di base.

R

R è un linguaggio di programmazione, tuttavia, il che significa che la curva di **apprendimento iniziale è più ripida**. Ci vorranno almeno alcune settimane per familiarizzare con l'interfaccia e padroneggiare le varie funzioni. Fortunatamente, l'uso di R può diventare rapidamente una seconda natura con la pratica.

Perchè programmare invece di usare Excel?

- R aiuta a leggere qualsiasi tipo di dato disponibile.
- L'automazione è molto più semplice rispetto a Excel.
- Supporta set di dati più grandi
- Offre anche un calcolo più veloce
- Le capacità di manipolazione dei dati sono potenti rispetto a Excel.
- Organizzazione del prodotto più semplice
- R può anche aiutare a rilevare qualsiasi tipo di errore (riproducibilità)
- È gratuito e open source. Quindi non è necessario pagare come Excel.
- R fornisce funzionalità statistiche avanzate.

Installazione e configurazione di R e RStudio

- https://vincnardelli.com/cbd/guide/installazione_r_rstudio
- https://vincnardelli.com/cbd/guide/usare_r_rstudio