

# Progetto SQL

di Matteo Francesco  
Biasio



Link al repository:



Grafici Tableau:



# Progetto SQL - premesse

- \* Nell'economia del progetto SQL, mi sono concentrato sullo sviluppo di uno script SQL atto a monitorare lo stato dell'arte al 2023 di alcune variabili, presenti nel DB «world-data-2023» **socio-economiche** , quali
  - \* **% di aree forestali**
  - \* **% di aree agricole**
  - \* **% lorda di iscrizioni a poli di istruzione terziaria**
  - \* **Aspettativa di vita**
  - \* **Tasso di mortalità infantile**
  - \* **Tasso di mortalità materna**
  - \* **GDP pro-capite**

# Progetto SQL - premesse

- \* Per quanto riguarda il DB «sustainableenergy» mi sono concentrato su
  - \* Produzione di energia da fonti rinnovabili (TWh)
    - Electricity generated from renewable sources (hydro, solar, wind, etc.) in terawatt-hours.
  - \* Percentuale di energia primaria derivata da fonti rinnovabili - Renewables (% equivalent primary energy)
  - \* Capacità di produzione di energia rinnovabile pro-capite

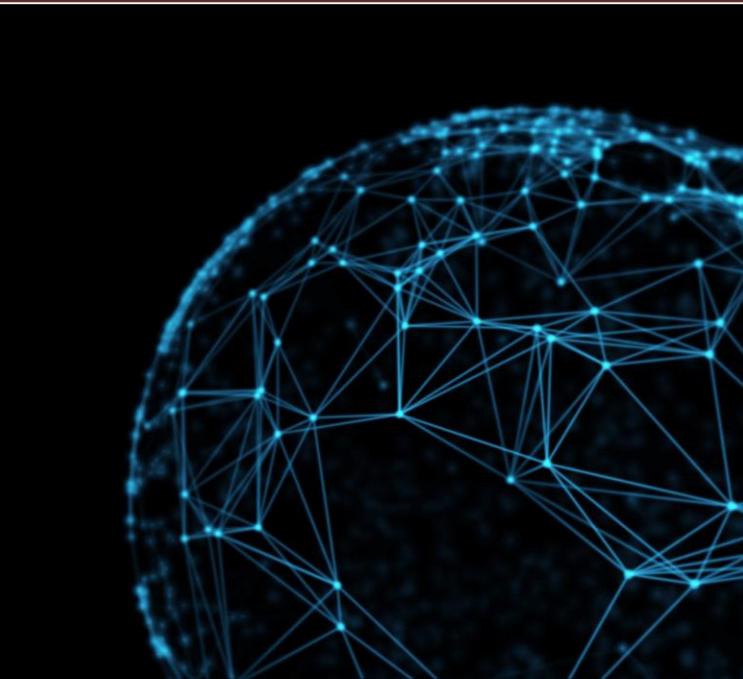


# Progetto SQL - premesse

La scelte appena citate sono motivate dal fatto di voler valutare molteplici *correlazioni* tra variabili, al fine di poter trarre insights significativi riguardo fenomeni socio-economici e monitorare quanto questi contesti, nel tempo, si siano rivelati «all'avanguardia» e «nuovi pionieri», per quanto concerne politiche atte alla produzione di energie rinnovabili.



# Creazione del DB - WordData2023



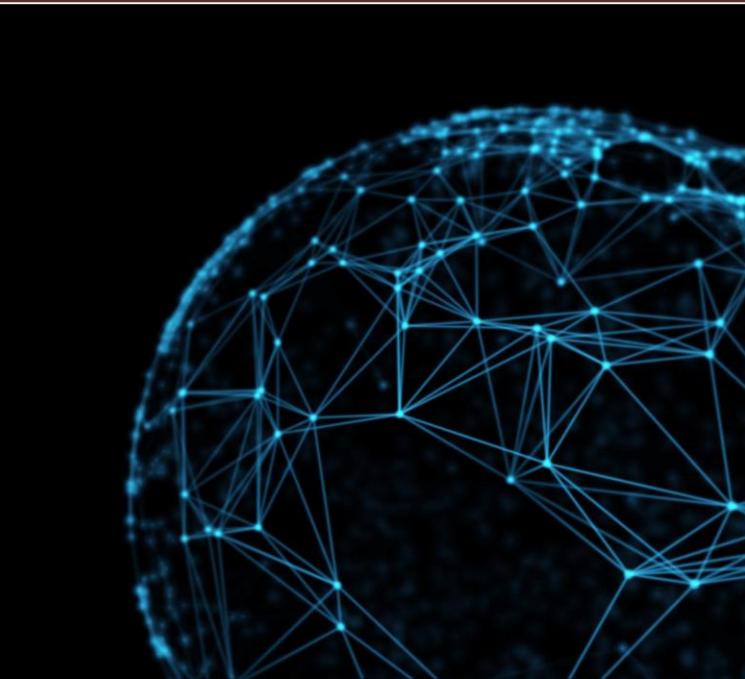
Le tipologie, la dimensione delle variabili e la non nullità sono state valutate in funzione del dataset originale. Le valutazioni possono essere monitorate in «world-data-2023\_cleaning\_procedure» .csv/.xlsx, in cui sono presenti i calcoli delle dimensioni delle stringhe con le funzioni disponibili in Excel, per ridurre lo spazio allocato nel DB [sfruttando annidamenti  $\text{MAX}(\text{LUNGHEZZA}(\text{colonna}))$ ], (usando poi su postgres una varchar - character varying) e dove ho valutato la dimensione delle variabili intere (smallint o int-integer). Inoltre:

- \* Ogni variabile intera è stata formattata come numero senza separatori delle migliaia, per risolvere i problemi in importazione del CSV su PostgreSQL (su STATA e Excel non davano mai problemi ["," come separatore dei campi]). Ho usato come separatori quelli che uso di sistema (derivazione anglosassone): punto per il decimale, virgola per le migliaia (in questo caso evitata per non dare problemi in fase di importazione).

# Creazione e popolazione del Dataset



# Creazione del DB - WordData2023



- \* Le percentuali, come spiegato, sono state riformattate in base "1" per il 100%, con un range per le cifre significative variabile, in funzione del tipo di dato proposto. Ho usato una variabile decimal-numeric.
- \* Estrazione del valore numerico della valuta usando un annidamento

*STRINGA.ESTRAI(testo, inizio,LUNGHEZZA(testo)).*

# Procedura di popolazione del Dataset

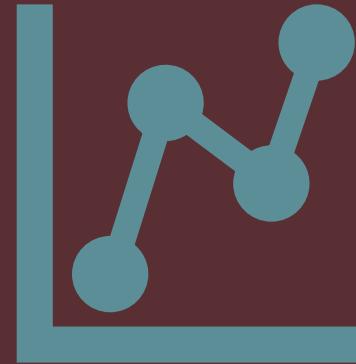
Come spiegato nello script SQL, ho adottato la strategia di importazione tramite il tool di PostgreSQL

- \* Tasto DX sulla tabella «worlddata2023»
- \* Import/Export Data...
- \* Selezionando correttamente "world-data-2023\_cleaned.csv" che mostra il DB ripulito e pronto all'importazione



# Correlazioni

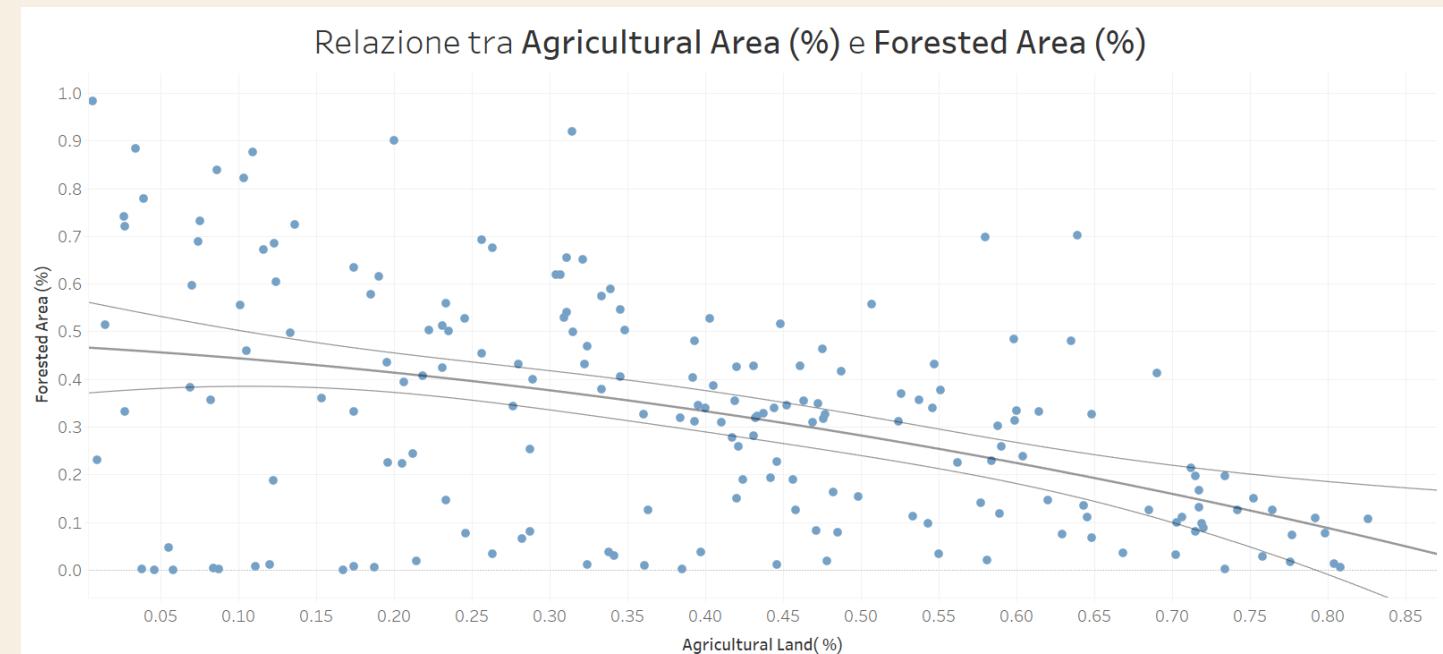
# WorldData2023 – principali analisi (correlazione)



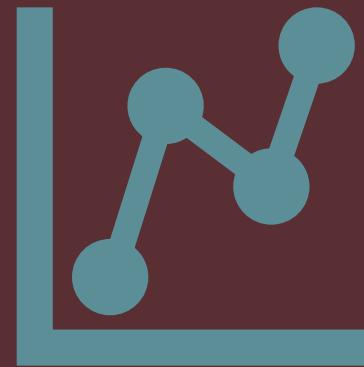
- \* Analizzando la correlazione  $\text{corr}(\text{var1}, \text{var2})$  tra la percentuale di terreno agricolo e la percentuale di terreno forestale è emersa una correlazione pari a

-0.434569556619751

La correlazione tra le due variabili è abbastanza banale e molto debole, ma può darci già una prima indicazione: la relazione "spaziale" sulla copertura territoriale indica una correlazione negativa: all'aumentare del tasso di terreno destinato ai fini agricoli, si riduce il territorio "incontaminato", coperto da foreste.

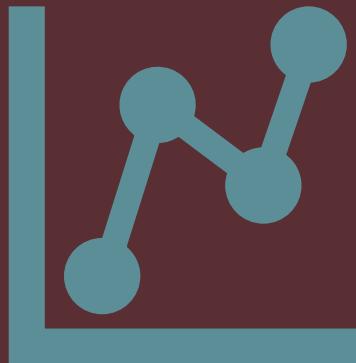


# WorldData2023 – principali analisi (correlazione)



- \* Rappresenta, in modo abbastanza agevole, il tipico fenomeno di «sottrazione» di territorio incontaminato ai fini del soddisfacimento del fabbisogno primario si sostentamento, attraverso la produzione agricola.  
Questa potrebbe avere un impatto ambientale non indifferente, soprattutto quando si parla di coltura intensiva, nella produzione di co2. L'effetto combinato della deforestazione può indurre al crearsi di un *effetto a catena*: la riduzione di copertura di foreste porta a non avere sufficiente quota «GREEN» per compensare la crescita di produzione di co2.  
MA il tema agricolo non necessariamente può spiegare complessivamente la questione GHG: occorrerebbe avere informazione anche circa la concentrazione di imprese ad alto tasso inquinante, così come informazioni sul tasso di concentrazione della popolazione.

# WorldData2023 – principali analisi (correlazione)

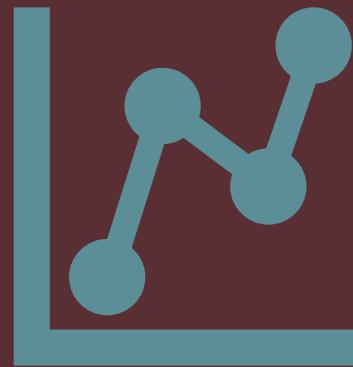


Un altro indicatore di interesse collettivo potrebbe essere rappresentato dall'aspettativa di vita «lifeexpectancy». Quest'ultima è fortemente collegata

- \* alle scelte che un individuo prende nell'economia della propria esistenza.
- \* Come probabilmente è anche collegata ai contesti economico-strutturali che le diverse nazioni propongono.

Sotto queste ipotesi è bene considerare come potenzialmente un individuo consapevole, con gli strumenti giusti, creandosi una rete di contatti adeguata, possa aspirare ad una *aspettativa di vita più longeva*.

# WorldData2023 – principali analisi (correlazione)

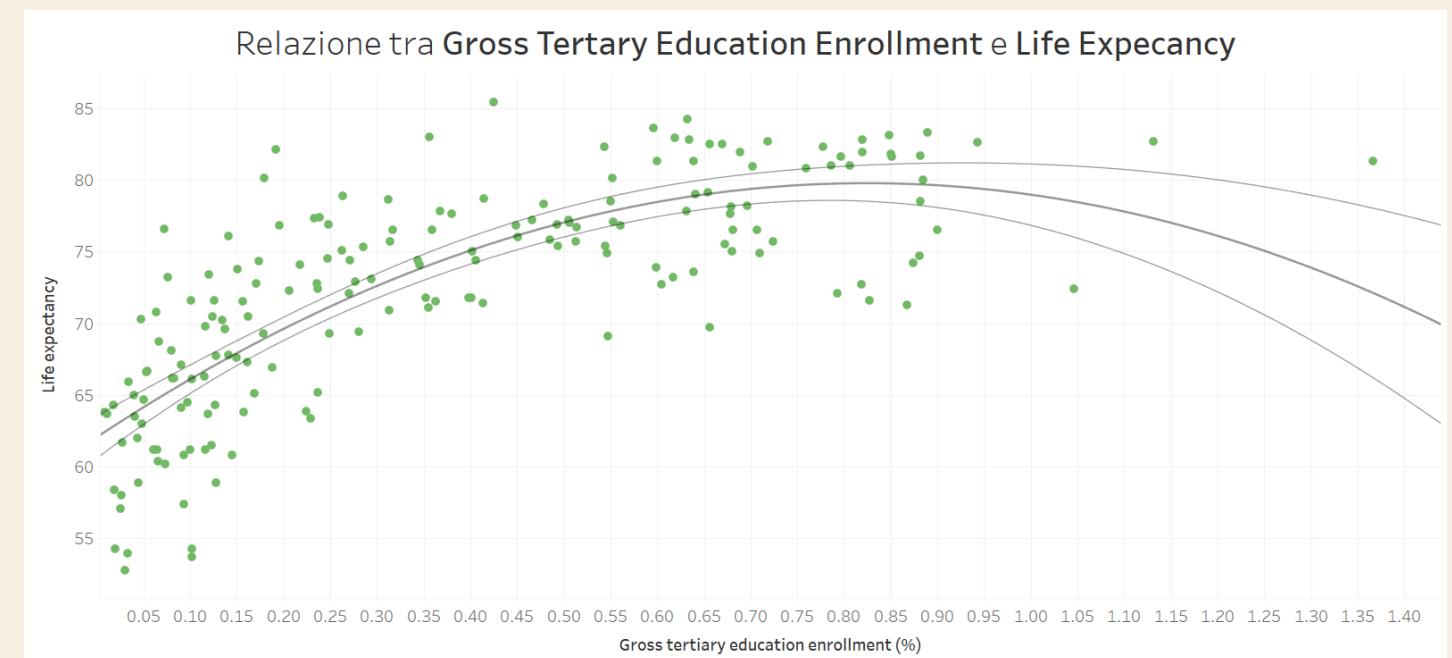


Per questo motivo ho provato ad analizzare una correlazione tra il

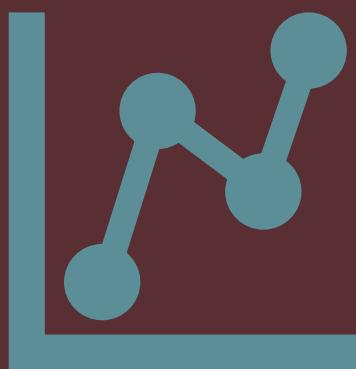
- \* tasso di partecipazione e iscrizione ad educazione terziaria «grosstertiaryeducationenrollmentperc» e
- \* l'aspettativa di vita «lifeexpectancy» (sopracitata).

L'output mostra una correlazione pari a:

0.7225347334725586



# WorldData2023 – principali analisi (correlazione)



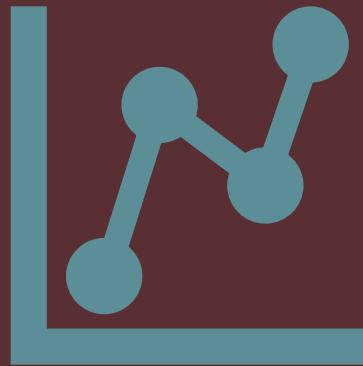
La cosa sarà sicuramente in parte motivata da una questione strutturale delle nazioni, anche dipendente dal livello della sanità. Ma, potenzialmente, la partecipazione a contesti accademici superiori-universitari può incidere fortemente sulle scelte del quotidiano:

- \* l'apertura mentale, la comunicazione con diverse culture, l'allargamento dello spettro conoscitivo che tali contesti possono offrire, mettono nelle condizioni, se disponibili ad "accettare *il nuovo*", a creare una maggiore consapevolezza nelle scelte del quotidiano (dalla scelta delle materie prime in un supermercato, fino all'aumento di propensione al rischio in fase di scelte finanziarie [investimenti, accesso ad un mutuo]).

Nell'economia comportamentale, il parametro istruzione gioca un ruolo fondamentale, in quanto permette di prendere scelte con più raziocinio e calma, riducendo lo stress. La componente scelta quotidiana e la conseguente riduzione dello stress possono incidere inevitabilmente sull'aumento di aspettativa di vita.

L'istruzione e il sapere rendono liberi, e la libertà incide sul benessere.

## WorldData2023 – principali analisi (correlazione)



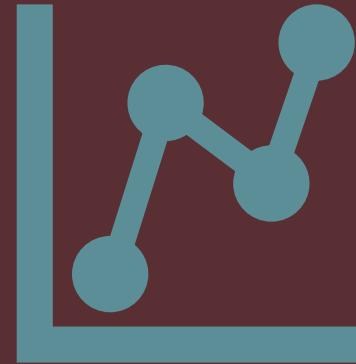
Data la premessa, si prosegue facendo un appunto sulla mortalità infantile/mortalità materna («*infantmortality*» (1) e «*maternal mortality*» (2)) e la correlazione con l'aspettativa di vita.

Quasi sicuramente queste coppie di variabili saranno autocorrelate fra loro, ma risulta comunque interessante valutarne l'entità:

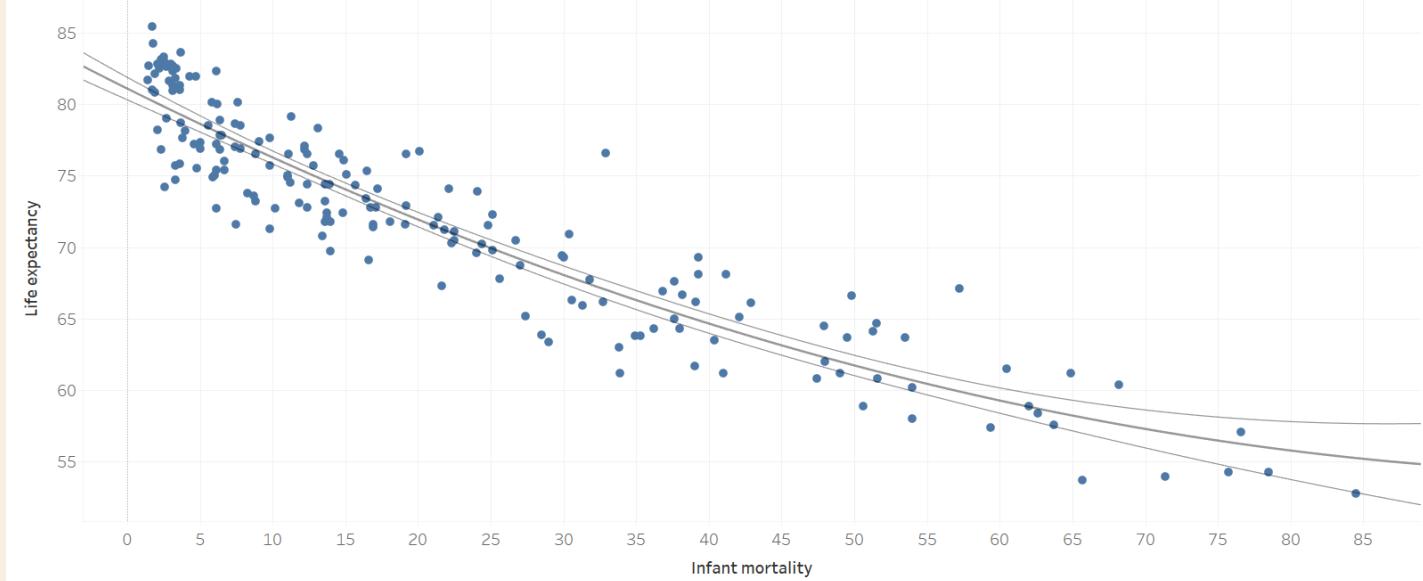
- (1): -0.9246753226745087
- (2): -0.8317965161194117

Risulta abbastanza lampante che la mortalità prematura e la mortalità materna si correlino, in qualche modo, con l'aspettativa di vita. Le opportunità e le condizioni di un sistema sanitario/individuali fragili incidono in modo pregante (negativamente) sull'aspettativa di vita.

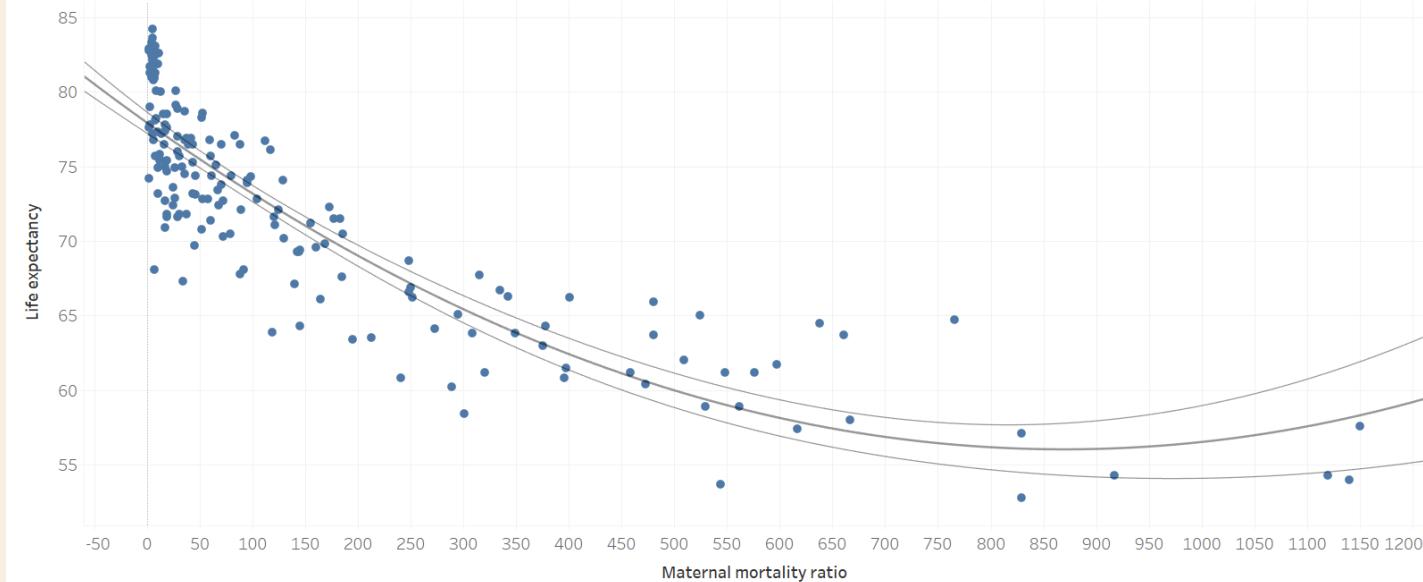
# WorldData2023 – principali analisi (correlazione)



Relazione tra Infant Mortality e Life Expectancy



Relazione tra Maternal Mortality Ratio e Life Expectancy

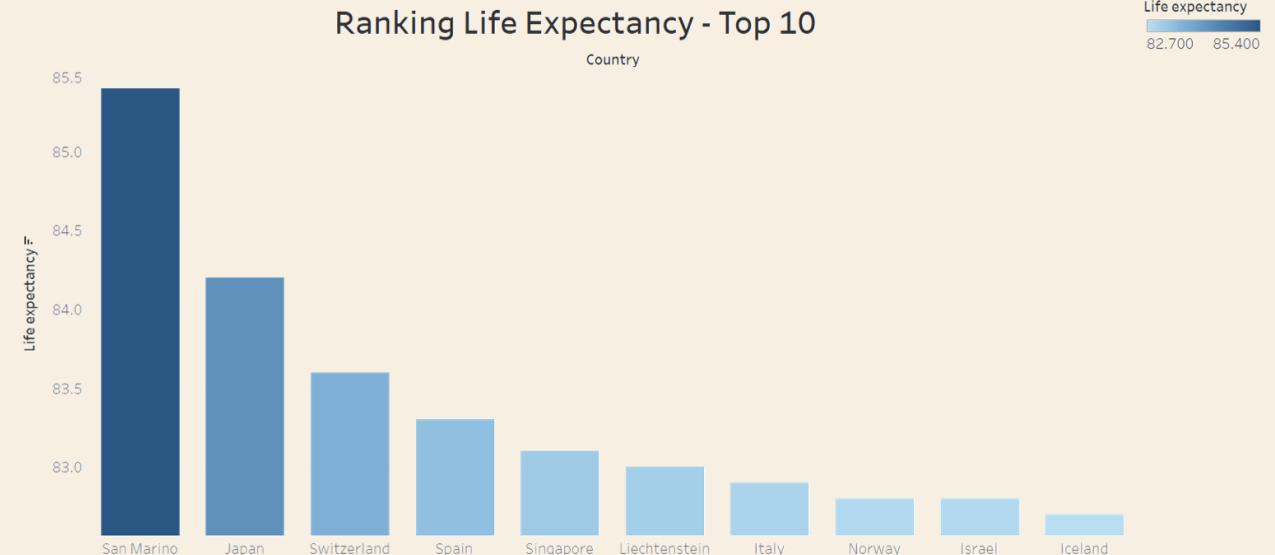


Ranking: Top & Worst

# Top 10: Life expectancy

- \* Come potrà mostrare la view «`top_life_exp`» le nazioni nella top 10 per aspettativa di vita sono in gran parte rappresentate da nazioni europee, fatta eccezione per il **Giappone** e **Singapore**.
- \* I casi **italiani** e **giapponesi** sono ormai noti da anni: sono tra le nazioni più anziane sul globo terracqueo. Nel caso italiano in più sappiamo quanto questo dato sia influenziato dall'alto **tasso di pensionati**, che non partecipato più direttamente alla produzione di GDP/PIL nazionale.

Country	Life Expectancy
San Marino	85.4
Japan	84.2
Switzerland	83.6
Spain	83.3
Singapore	83.1
Liechtenstein	83.0
Italy	82.9
Norway	82.8
Israel	82.8
Iceland	82.7



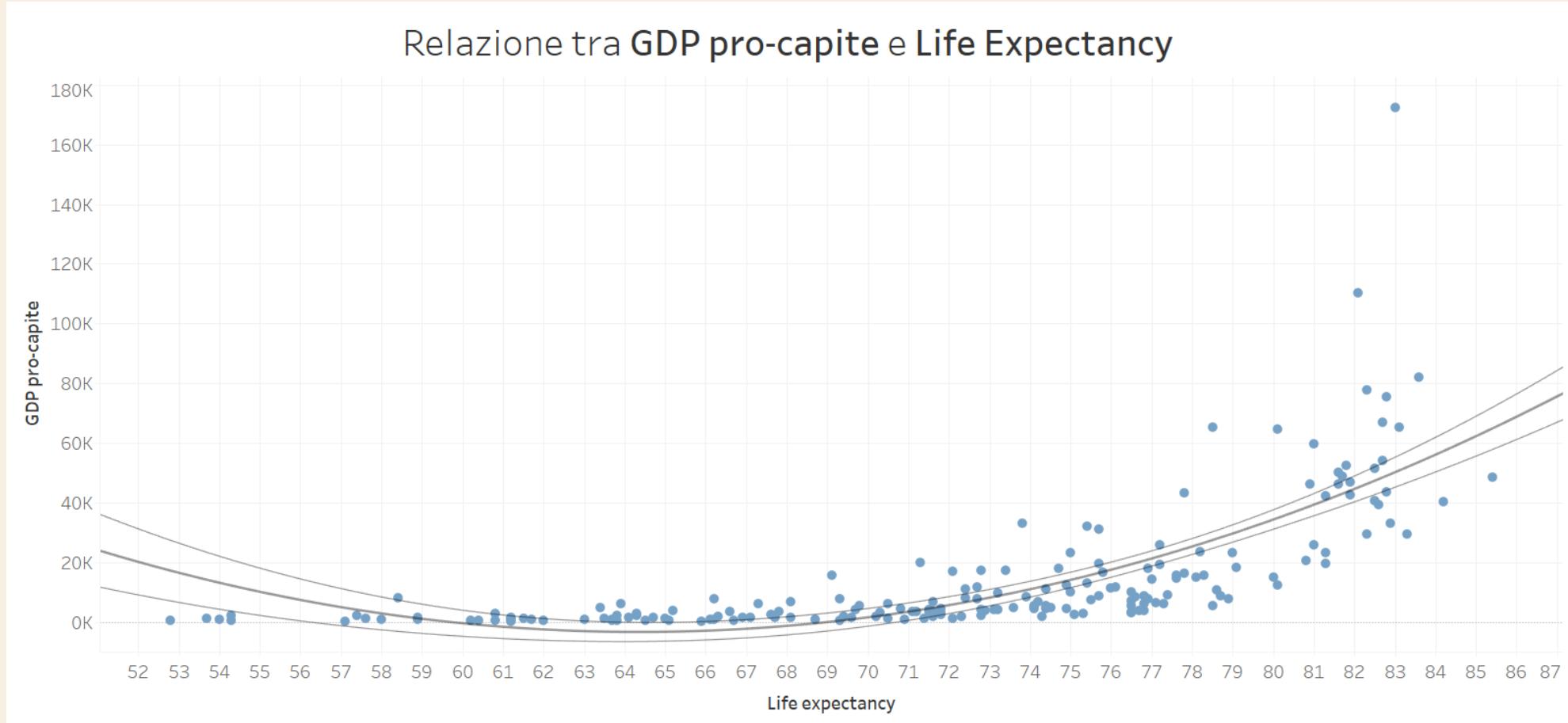
# Top 10: Life expectancy - correlazioni

- \* Dunque, servirebbe dare un'occhiata a se la **produzione di benessere per la collettività** sia in qualche modo **correlata con l'aspettativa di vita**. Prendendo in esame il **gdp pro-capite** («GDP»/»population») e «lifeexpectancy» si potrebbe evidenziare, attraverso una analisi di **correlazione (3)**, che le nazioni con più alta produzione di Prodotto Interno Lordo - pro-capite - si troverebbero mediamente in una situazione strutturale tale da garantire un sistema "salubre", favorendo la crescita dell'aspettativa di vita.

(3) 0.6150761448655165

L'indagine a livello grafico, tuttavia, ci porta ad evidenziare una significatività debole della correlazione.

# Top 10: Life expectancy - correlazioni



# Worst 10: Life expectancy

- \* La view «`worst_life_exp`» evidenzia che la situazione peggiore per la «`lifeexpectancy`» la si può osservare invece nei paesi **centrafricani equatoriali**

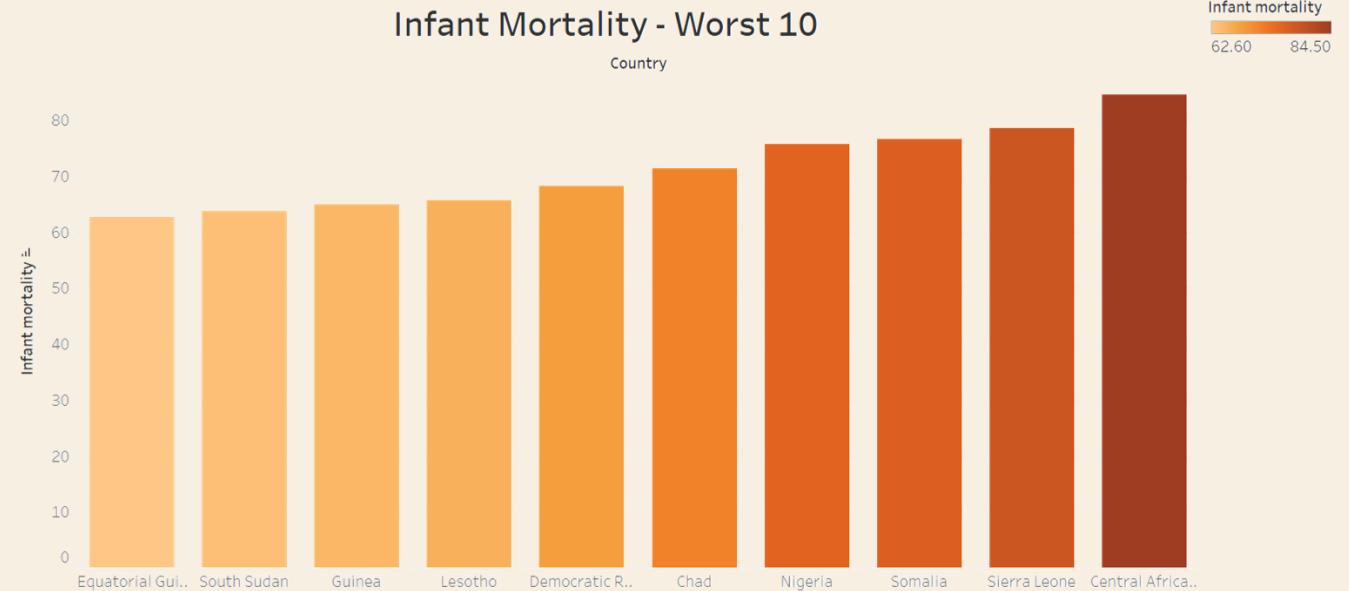
Country	Life Expectancy
Central African Republic	52.8
Lesotho	53.7
Chad	54.0
Sierra Leone	54.3
Nigeria	54.3
Somalia	57.1
Ivory Coast	57.4
South Sudan	57.6
Guinea-Bissau	58.0
Equatorial Guinea	58.4



# Worst 10: Infant mortality

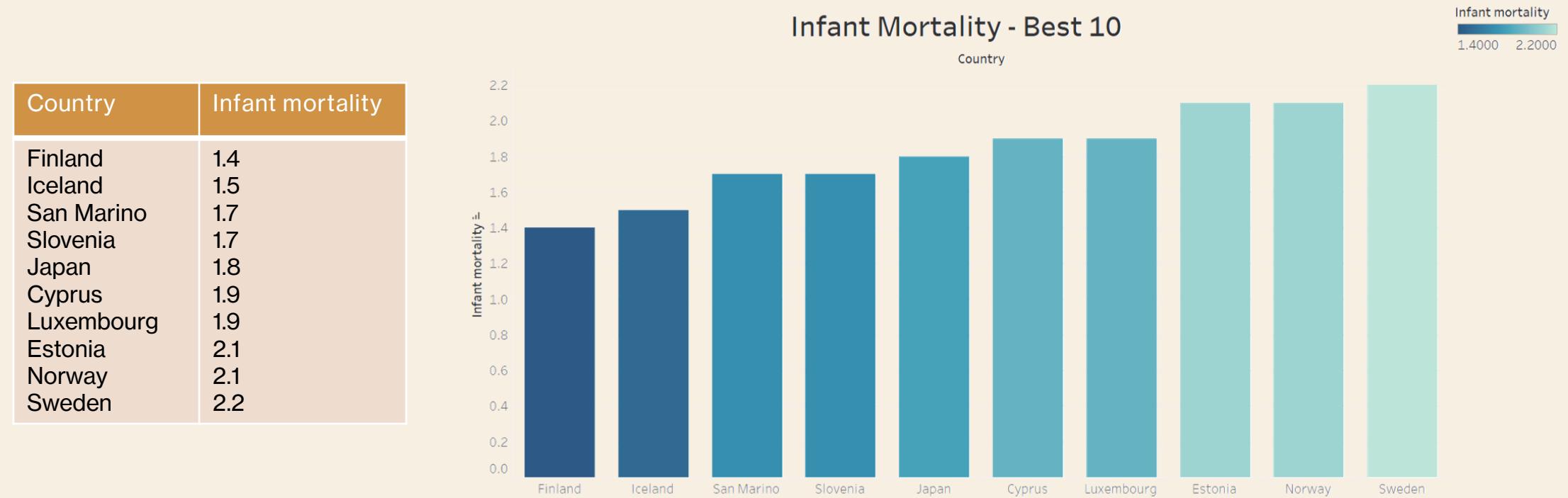
- \* La view «*worst\_infant\_mortality*» invece suggerisce che il livello più alto di mortalità infantile «*infantmortality*» è osservabile in quelle nazioni il cui contesto sanitario è debole, talvolta quasi assente, come nei **territori centrafricani**

Country	Infant mortality
Central African Republic	84.5
Sierra Leone	78.5
Somalia	76.6
Nigeria	75.7
Chad	71.4
Democratic Republic of the Congo	68.2
Lesotho	65.7
Guinea	64.9
South Sudan	63.7
Equatorial Guinea	62.6



# Top 10: Infant mortality

- Nella view «**best\_infant\_mortality**», all'altro lato della distribuzione, osserviamo la grande maggioranza di **nazioni europee**, fatta eccezione per il **Giappone**. Non compare l'Italia ai primi posti, ma solo al 14esimo.



# Worst 10: Maternal mortality

- \* Con il supporto della view «**worst\_maternal\_mortality**» si evince che il livello più alto di mortalità materna è osservabile in quelle nazioni il cui contesto sanitario è debole, tavolta quasi assente, come nei territori **centrafricani** o in **Afghanistan**.

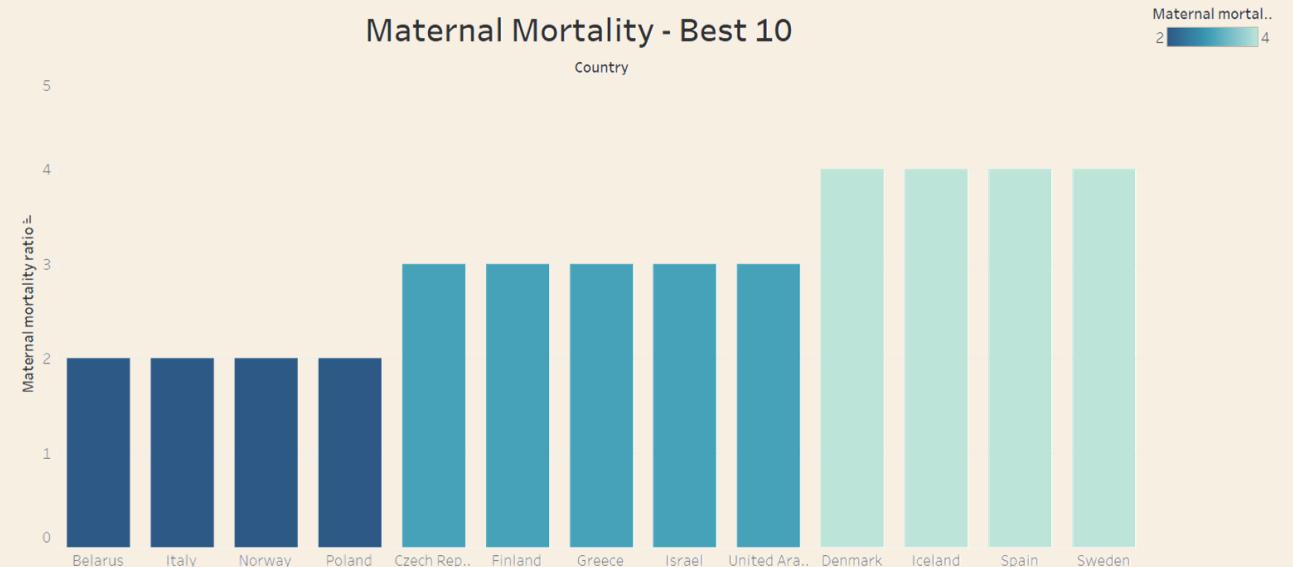
Country	Maternal mortality ratio
South Sudan	1150
Chad	1140
Sierra Leone	1120
Nigeria	917
Central African Republic	829
Somalia	829
Mauritania	766
Guinea-Bissau	667
Liberia	661
Afghanistan	638



# Top 10: Maternal mortality

- Come potrà mostrare la view «**best\_maternal\_mortality**» all'altro lato della distribuzione osserviamo la grande maggioranza di nazioni europee, fatta eccezione per gli **EAU**.

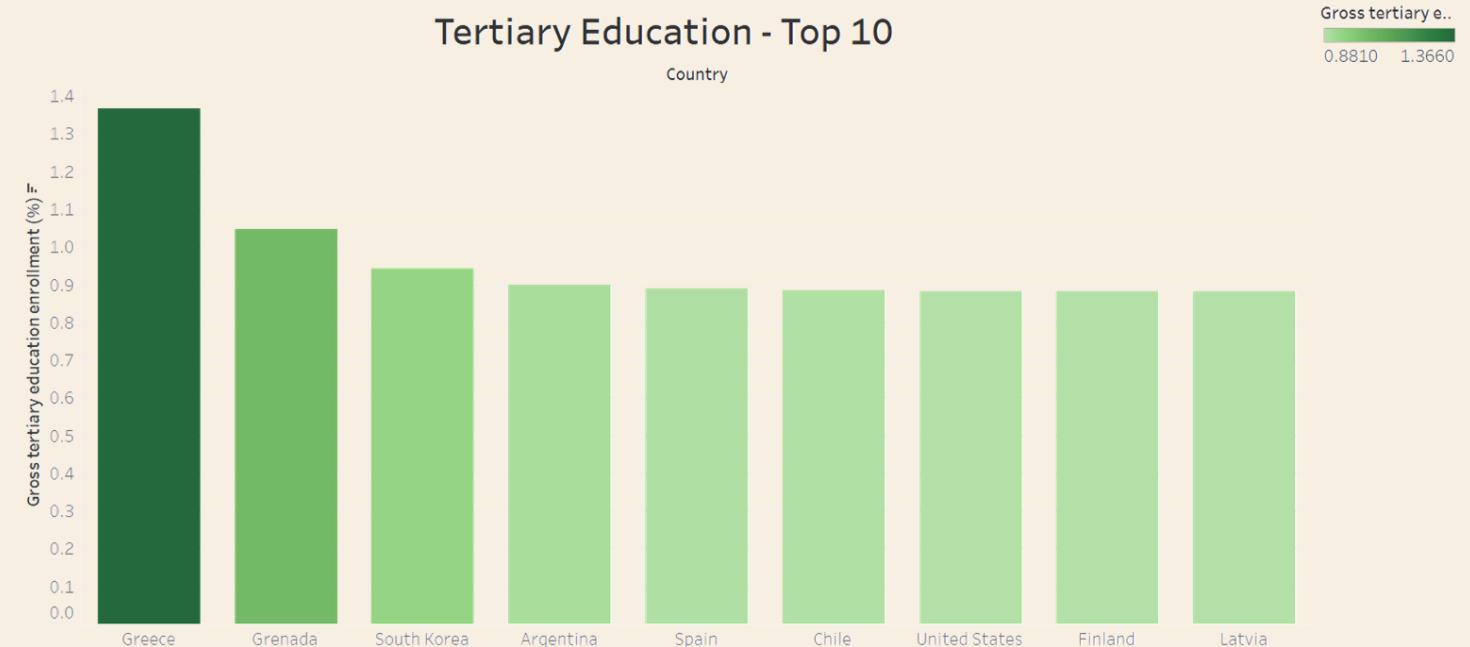
Country	Maternal mortality ratio
Poland	2
Norway	2
Belarus	2
Italy	2
Finland	3
Greece	3
Czech Republic	3
Israel	3
United Arab Emirates	3
Iceland	4



# Top 10: Tertiary education

- Nella view «**top\_ter\_edu**» il livello educativo, in termini di grado di iscrizione linda, vede una distribuzione alquanto differenziata fra le varie parti nel mondo: al lordo di potenziali studenti e cittadini stranieri, troviamo la **Grecia** in testa, seguita da **dall'Australia, Grenada, South Korea, Argentina, Spagna...**

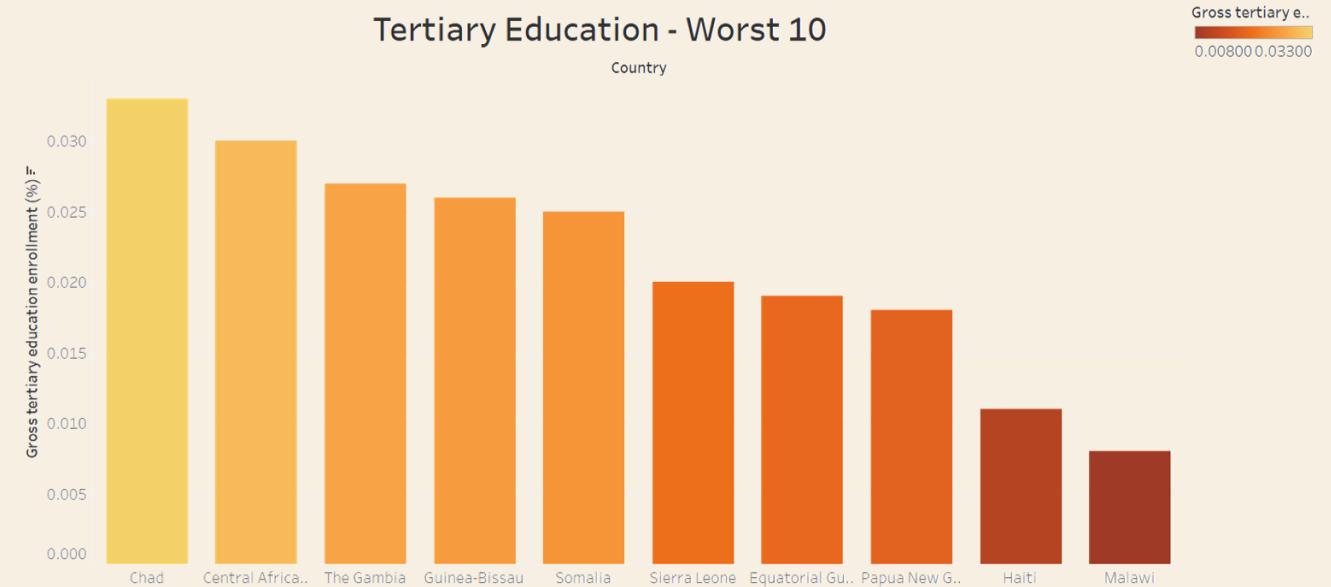
Country	Gross tertiary education enrollment percentage
Greece	1.3660
Australia	1.1310
Grenada	1.0460
South Korea	0.9430
Argentina	0.9000
Spain	0.8890
Chile	0.8850
Finland	0.8820
United States	0.8820
Latvia	0.8810



# Worst 10: Tertiary education

- \* La view «`top_ter_edu`» si riferisce al livello educativo e suggerisce, come prevedibile, le nazioni con il più basso tasso di iscrizione a percorsi di educazione terziaria sono **centrafricane**, fatta eccezione per **Haiti**

Country	Gross tertiary education enrollment percentage
Malawi	0.0080
Haiti	0.0110
Papua New Guinea	0.0180
Equatorial Guinea	0.0190
Sierra Leone	0.0200
Somalia	0.0250
Guinea-Bissau	0.0260
The Gambia	0.0270
Central African Republic	0.0300
Chad	0.0330



# Top 10: Life expectancy and Tertiary education

- \* Riprendendo il discorso sulla correlazione positiva tra la **life expectancy** e la **percentuale di iscrizione a percorsi di educazione terziaria**, possiamo vedere se vi sono nazioni che risultano nella top 10 di entrambi gli indicatori.
- \* La view «**top\_life\_tertiaryED**» mostra come solo la **Spagna** si trovi nella TOP 10 come **aspettativa di vita** e come **percentuale lorda di iscrizione a percorsi di formazione terziaria**

Country	Life expectancy	Gross tertiary education enrollment percentage
Spain	83.3	0.8890

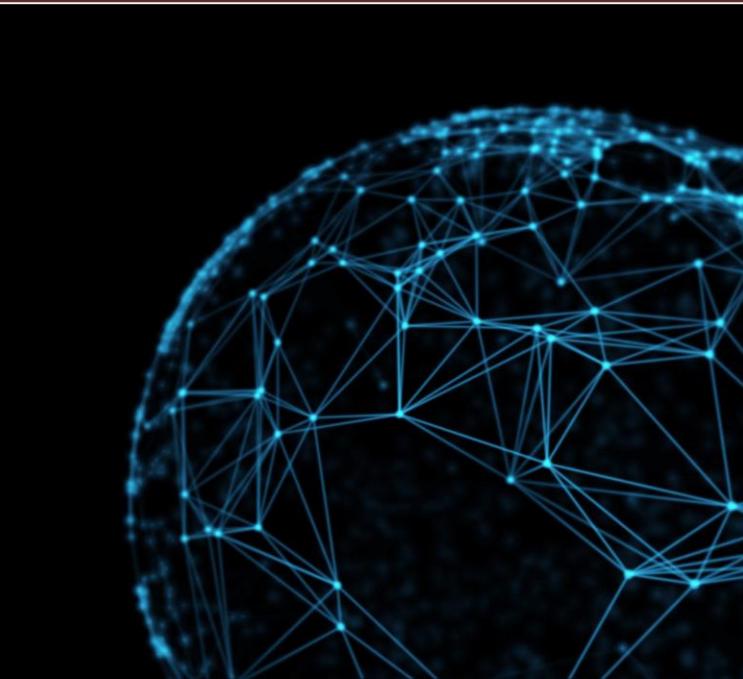
# Worst 10: Life expectancy and Tertiary education

- \* Analizzando, invece, la vista «**worst\_life\_tertiaryED**» in questa lista viene mostrata una situazione alquanto prevista e preoccupante: su 10 delle precedenti due liste, 6 nazioni risultano tra le peggiori situazioni sia in termini di **istruzione** che di **aspettativa di vita** e sono tutte del contesto **centroafricano**.

Country	Life expectancy	Gross tertiary education enrollment percentage
Equatorial Guinea	58.4	0.0190
Sierra Leone	54.3	0.0200
Somalia	57.1	0.0250
Guinea-Bissau	58.0	0.0260
Central African Republic	52.8	0.0300
Chad	54.0	0.0330

# Sustainable Energy Dataset

# Creazione del DB – sustainableenergy



Per le ragioni già motivate per il dataset WorldData2023, sono state fatti degli adattamenti sul formato dei dati sorgente:

- \* le modifiche sono osservabili nei file "global-data-on-sustainable-energy (1)\_cleaning\_procedure« in .csv/.xlsx
- \* Per le stringhe viene identificata la lunghezza massima e utilizzato il tipo di dato VARCHAR - CHARACTER VARYING, in modo da rendere adattivo lo spazio allocato su disco
- \* per le percentuali viene adottata una trasformazione in base 1 per il 100%: vengono identificate a monte le cifre significative ed adattate in funzione della trasformazione. Formato DECIMAL/NUMERIC Y cifre significative. (PostgreSQL converte automaticamente il formato)
- \* per i decimal/float/numeric si identificano le lunghezze massime delle stringhe SENZA pointer, valutato il MAX e il MIN della serie e valutate le cifre significative da adottare
- \* gli interi vengono valutati in funzione della dimensione massima del numero (allocando smallint-int-bigint in base alla necessità)
- \* rappresentazione dei dati della densità di popolazione senza separatori delle migliaia ("," dava problemi in fase di importazione, legge tale carattere come separatore di valore, sempre per la questione dello standard anglosassone che uso sul PC).

## Procedura di popolazione del Dataset



Come spiegato nello script SQL, ho adottato la strategia di importazione tramite il tool di PostgreSQL

- \* Tasto DX sulla tabella "sustainableenergy"
- \* Import/Export Data...
- \* Selezionando correttamente "global-data-on-sustainable-energy (1)\_cleaned.csv" che mostra il DB ripulito e pronto all'importazione, originato da "global-data-on-sustainable-energy (1).csv" e che vede le modifiche in "global-data-on-sustainable-energy (1)\_cleaning\_procedure.csv" e "global-data-on-sustainable-energy (1)\_cleaning\_procedure.xlsx"

# JOIN World Data & sustainable energy datasets



# JOIN WordaData2023 – sustainableenergy e metriche di performance



Passaggio interessante è quello relativo al join del dataset WorldData2023 (nella versione a vista EnvEcoWorld2023) con sustainableenergy:

- \* l'obiettivo è avere un dataset unico, con chiave primaria rappresentata da
  - \* Country
  - \* annola granularità è su base annuale.
- \* per monitorare agilmente alcune metriche negli anni ho pensato di sfruttare un *monitoraggio*, generando la media mondiale annuale di specifici indicatori, *dei trend nel tempo* di
  - \* produzioni nazionali di Twh di elettricità con fonti rinnovabili «elecrenew\_twh»

# JOIN WordaData2023 – sustainableenergy e metriche di performance



- \* Vengono invece mantenute invariate
  - \* La capacità di produzione di energia rinnovabile pro-capite
  - \* percentuale di energia primaria derivata da fonti rinnovabili «Renewables (% equivalent primary energy)»
- \* Le prima metrica nazionale, per anno, viene rimodulata in ragione della MEDIA MONDIALE ANNUALE. Quindi questa verrà intesa come METRICA ANNUALE RELATIVA
  - \* > 1 se la metrica nazionale annuale rimane sopra il valore di SOGLIA della media mondiale annuale
  - \* = 1 se la metrica nazionale annuale rispecchia quella mondiale
  - \* < 1 se la metrica nazionale annuale è sotto il valore di SOGLIA della media mondiale annuale

$$national\ realative\ METRIC\ YYYY_j = \frac{national\ metric\ YYYY_j}{world\ average\ metric\ YYYY}$$

$j = 1, \dots, N$   $j$  – esima nazione  
 $YYYY$  2000 ... 2020 anno di misurazione

# Sustainable energy – time-series analysis

\*Monitorata la compresenza dello stesso indicatore di «landarea\_km2» ed eliminando, dunque, il doppione, ho creato, dopo il join tra i due dataset, una vista «EnvEco2023\_SusEne» che permetta il monitoraggio degli indicatori relativi all'energia da fonti sostenibili.

\*Come si potrà osservare dallo script, ho creato la media annuale mondiale dell'indicatore

\* «elecrenew\_twh»



# Sustainable energy – time-series analysis

Vengono mantenute intatte

- \* «renewable-electricity-generating-capacity-per-capita»
- \* «Renewables (% equivalent primary energy)»  
(analizzata poi con Tableau)

\*Nel primo caso, questo valore viene usato come riferimento per rimodulare le misurazioni nazionali annuali in funzione di un valore soglia, rappresentato da questa media – creando così “elecrenew\_relative”. Lo strumento permette di osservare, anche con l’ausilio di un grafico, eventuali convergenze o divergenze dell’indicatore rispetto alla media globale.



# Sustainable energy – time- series analysis



- \*Tenuto conto della misurazione appena creata, richiamo la funzione *rank()* per mappare un ordinamento sulla base di questo indicatore, partizionando per anno.
- \*In questo modo ho creato una classifica annuale dell'indicatore, per mostrare quali sono le nazioni più performanti.

# Electricity from renewables (TWh) – time-series ranking

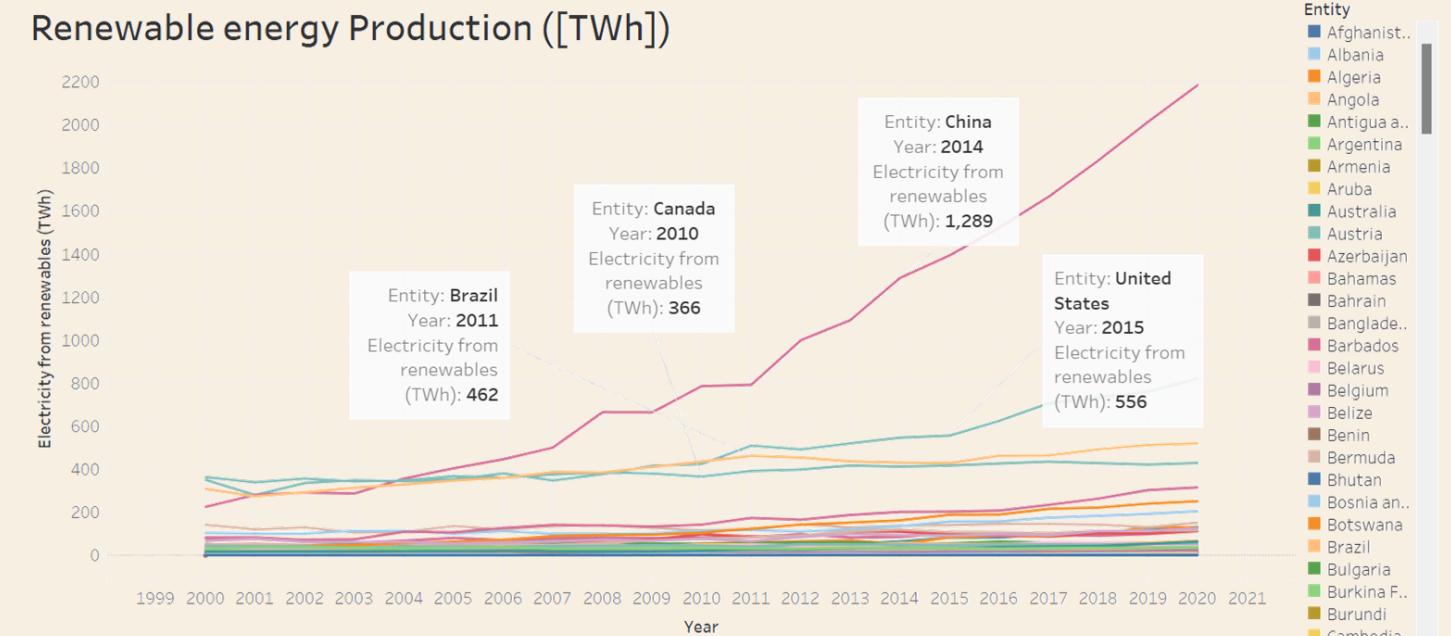


Come è possibile osservare dalla view «elec\_renew\_twh»,

- \* la Cina si mostra fra le prime nazioni per generazione di energia rinnovabile come Twh. Presenta una media largamente superiore alla media annuale mondiale «world\_annual\_average», con valori di produzione che passano dalle 15 volte la media mondiale al 2000 a 45-50 volte arrivando ai giorni nostri. Rispetto ai primi anni 2000, nel 2005 assume una posizione di leadership nella produzione (rank = 1) e la mantiene fino al 2020. Rispetto alla media mondiale, la produzione tenderebbe ad avere un andamento divergente.

Qui sotto il grafico del trend temporale ideato con Tableau (valori assoluti - TWh).

Renewable energy Production ([TWh])



# Electricity from renewables (TWh) – time-series ranking

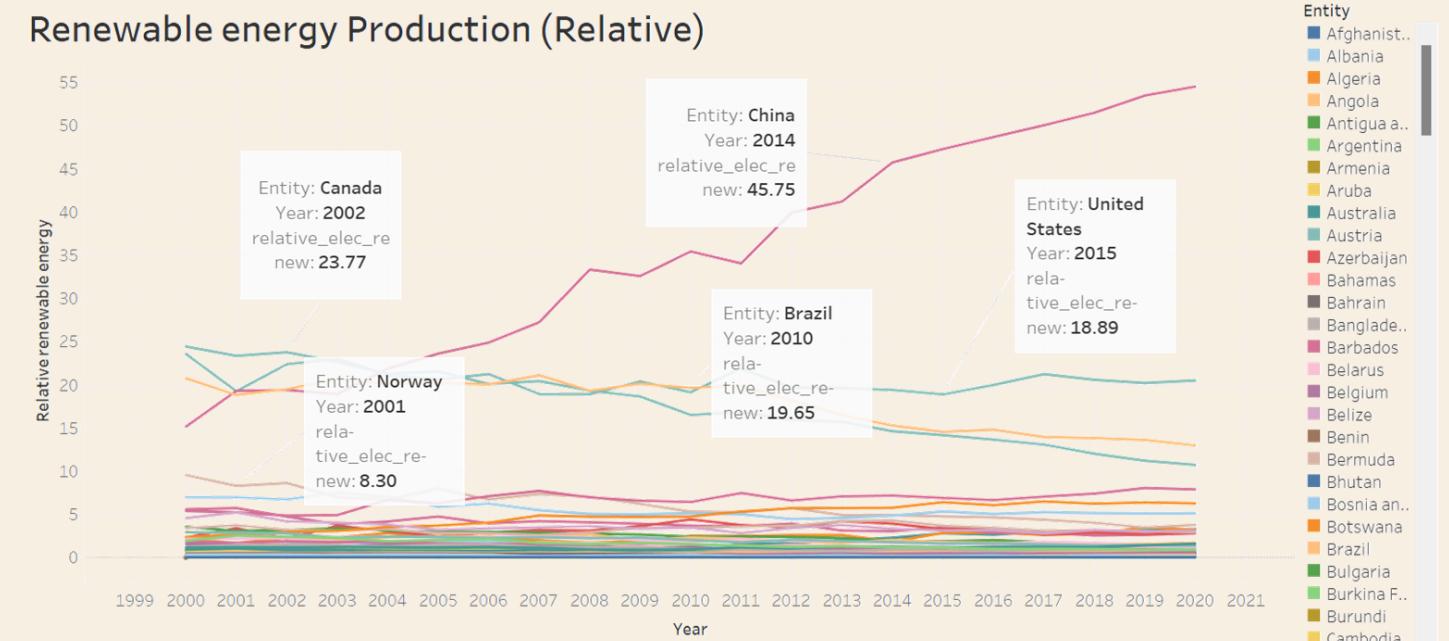


Come è possibile osservare dalla view «elec\_renew\_twh»,

- \* la Cina si mostra fra le prime nazioni per generazione di energia rinnovabile come Twh. Presenta una media largamente superiore alla media annuale mondiale «world\_annual\_average», con valori di produzione che passano dalle 15 volte la media mondiale al 2000 a 45-50 volte arrivando ai giorni nostri. Rispetto ai primi anni 2000, nel 2005 assume una posizione di leadership nella produzione (rank = 1) e la mantiene fino al 2020. Rispetto alla media mondiale, la produzione tenderebbe ad avere un andamento divergente.

Qui sotto il grafico del trend temporale ideato con Tableau (valori relativi).

Renewable energy Production (Relative)

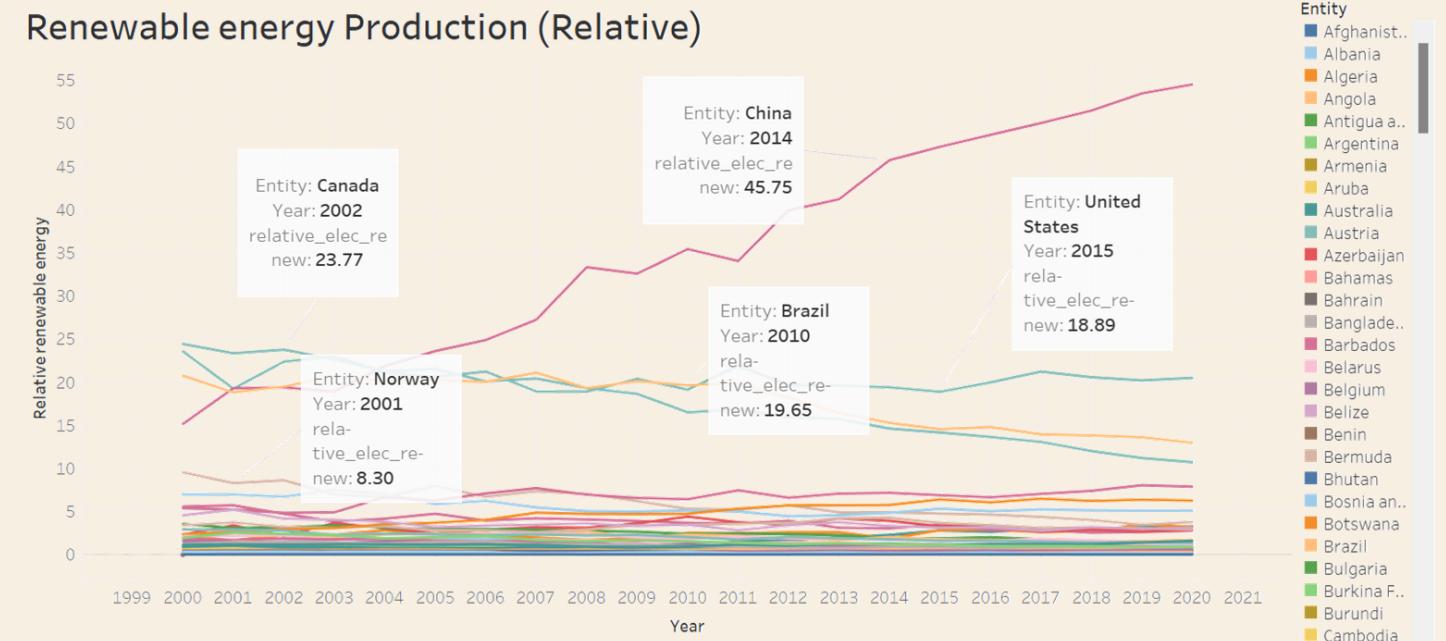


# Electricity from renewables (TWh) – time-series ranking



- \* Nel periodo 2000-2020 possiamo osservare come, in larga parte, le posizioni di testa, rispetto alla media mondiale annuale, siano occupate dal blocco Canada-Stati Uniti-Brasile-Cina-Norvegia, con India e Giappone che si fanno strada nel tempo.

Renewable energy Production (Relative)



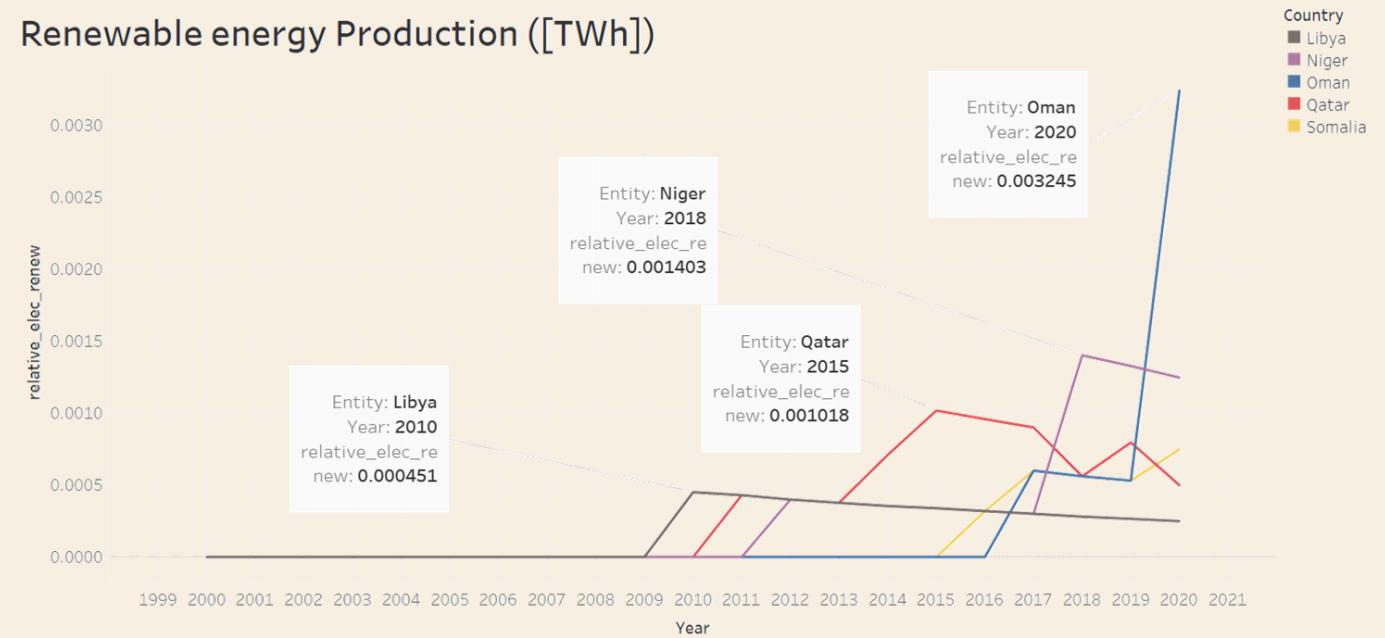
# Electricity from renewables (TWh) – time-series ranking



Per quanto riguarda le posizioni in fondo alla classifica (escludendo nazioni con campo vuoto o pari a 0) possiamo trovare, nel tempo, realtà depresse o che potrebbero aver iniziato progetti di conversione solo negli ultimi anni, quali:

- \* Oman, Niger, Somalia, Qatar, Libya

Renewable energy Production ([TWh])

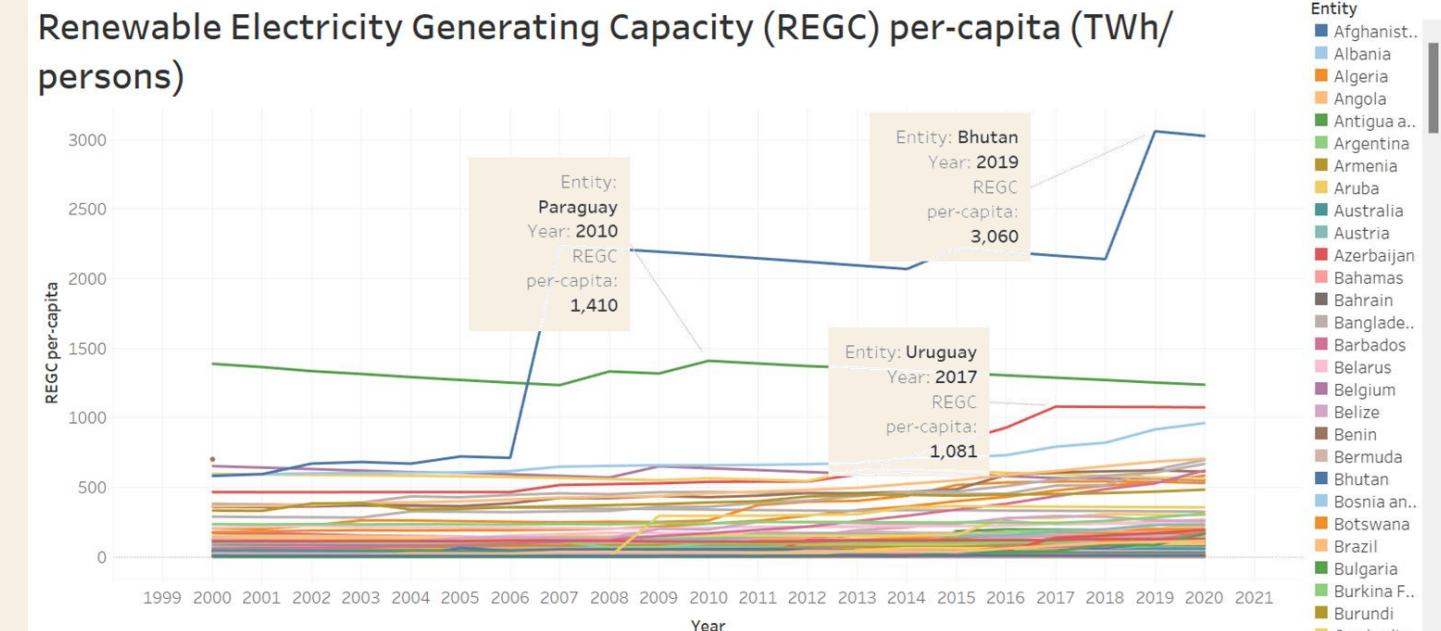


# Renewable Electricity Generating Capacity (REGC) per-capita (TWh/persons) – time-series ranking



Indagando ulteriormente, desta particolare interesse il riferimento alla capacità di produzione pro-capite di energia rinnovabile:

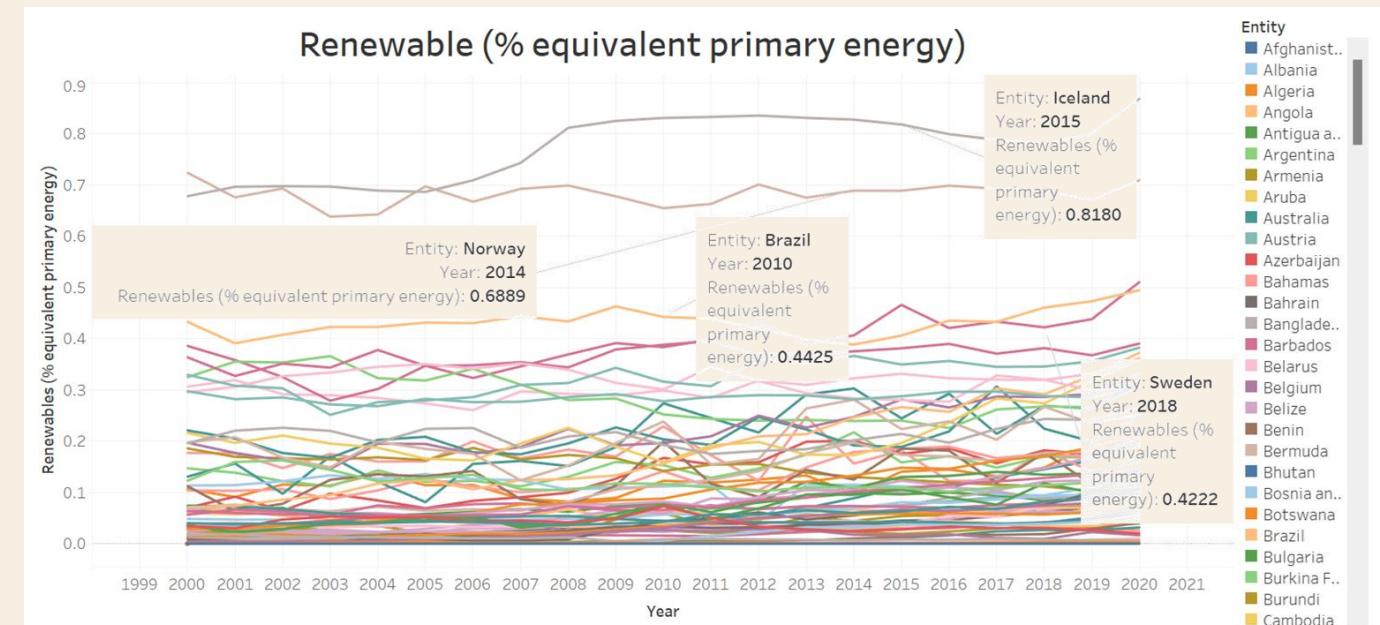
- \* Nazioni come il Bhutan e compagni sudamericane come Paraguay e Uruguay mostrano come producano una quantità di TWh pro-capite annua di gran lunga superiore rispetto a quella che sarebbe richiesta per un fabbisogno primario.
- \* In modo particolare, rispetto alla popolazione residente in Bhutan, la produzione pro-capite nazionale, nel tempo, supererebbe di gran lunga il fabbisogno dei cittadini locali.
- \* Buona parte dell'energia prodotta potrebbe essere indotta ad esportazione.



# Renewable energy share in total final energy consumption (%) – time-series ranking



Valutando, infine, la **percentuale di energia primaria equivalente**, si stima come i **paesi scandinavi**, accompagnati dal **Brasile**, si siano indirizzati nel tempo all'utilizzo di energia non soggetta a trasformazioni (energia primaria) di origine principalmente rinnovabile. In relazione a ciò, sembrerebbe lampante che queste realtà stiano adottando politiche volte ad un **mix energetico interno con quota di energia «green» sempre maggiore**.



# Considerazioni finali

# Considerazioni finali

In relazione a quanto emerso,

- \* appurate le dimensioni delle correlazioni tra molteplici variabili strutturali, che concorrono a motivare ed evidenziare lo stato di salute delle nazioni e
- \* Appurato come alcune nazioni, di specifici contesti, si trovino in condizioni *sanitarie, educative, ambientali, economiche* più o meno disagiate

Occorre far presente quanto alcuni contesti nazionali siano *proiettati* verso politiche *green* più accentuate rispetto ad altri - vedasi l'ultimo grafico della percentuale di energia primaria -.

## Considerazioni finali

Queste nazioni, in particolare le scandinave, risultano tra le nazioni con il più basso tasso di mortalità infantile e materno. Certo, questo potrebbe essere in qualche modo dipendente dalle politiche sanitarie e dalla qualità del servizio offerto.

Ma ci potrebbe essere ragione di dubitare riguardo al fatto che il fattore «ambientale» - *quello che respiriamo* -, insieme allo stile di vita, l'educazione, il contesto economico, non concorra a spiegare la mortalità.

La conversione di queste nazioni, negli ultimi vent'anni, a politiche energetiche rinnovabili e a meno impatto ambientale, potrebbe aver concorso, almeno in parte, a spiegare uno stato di salute individuale che *minimizzi le probabilità di morte prematura?*