Progetto Data Management

L’impronta invisibile: un’analisi delle emissioni di CO2 nel mondo

**Alessandro Bosi, Denis Bugaenco, Eleonora Zullo**

**Abstract**

L'emissione di CO2 ha un ruolo significativo e negativo sull'ambiente: è uno dei principali gas serra che intrappola il calore nell'atmosfera terrestre. Questo processo, noto come effetto serra, è essenziale per mantenere la Terra abbastanza calda da sostenere la vita. Tuttavia, l'aumento delle emissioni ha intensificato questo effetto, portando a un riscaldamento globale.

Ma quali sono i paesi che hanno contribuito maggiormente all’aumento dei livelli di anidride carbonica? Quanto è cambiata l’emissione negli anni? Ma soprattutto, quanto è incisivo il ruolo della popolazione nella produzione di questo gas?

Questo progetto esplora la relazione tra emissioni di CO2, densità di popolazione ed estensione territoriale in diversi paesi. Tramite l’analisi approfondita di due dataset, uno contenente dati su popolazione e superficie territoriale e l'altro riportante la produzione di CO2 per paese, la nostra analisi punta a rivelare correlazioni significative tra densità di popolazione ed emissioni di CO2, così come tra estensione territoriale e livelli di emissioni.

**Sommario**

[1. Data acquisition 1](#_Toc170927529)

[2. Data preparation 2](#_Toc170927530)

[2.1 Data cleansing 2](#_Toc170927531)

[2.2 Data integration 2](#_Toc170927532)

[3. Data storage 3](#_Toc170927533)

[4. Data quality 4](#_Toc170927534)

[5. Data analysis 4](#_Toc170927535)

[Conclusioni 5](#_Toc170927536)

[Ruolo dei partecipanti 5](#_Toc170927537)

[Riferimenti 6](#_Toc170927538)

## **Data acquisition**

Per l’acquisizione dei dati utilizzati per rispondere alle domande di ricerca del progetto, ci siamo proposti di lavorare su due dataset differenti.

Il primo dataset è stato ottenuto tramite un’Application Programming Interface dal seguente endpoint *https://restcountries.com/v3.1/all* e successivamente salvato nella collezione *countries\_collection* su MongoDB, contenuta nel database di riferimento denominato *countries\_db*.

Il dataset contiene 250 documenti, che rappresentano 250 stati differenti, ognuno dei quali è descritto dalle seguenti 10 variabili:

* *\_id*: il codice identificativo del paese,
* *name*: il nome del paese,
* *population*: la popolazione del paese,
* *region*: il continente di riferimento,
* *subregion*: l’area del continente in cui si trova il paese di riferimento,
* *capital*: la capitale del paese,
* *area*: l’estensione del paese in km2,
* *languages*: le lingue parlate nel paese,
* *borders*: gli stati confinanti con il paese,
* *currencies:* la valuta utilizzata nel paese.

Il secondo dataset, identificato come *pollution*, è stato ottenuto da Data.WorldBank, una piattaforma rinomata per la condivisione di dati di qualità e affidabilità relativi a vari indicatori economici, sociali e ambientali. Il dataset selezionato per il nostro progetto contiene informazioni dettagliate sulle emissioni di anidride carbonica pro capite per 266 paesi ed è stato ottenuto tramite download dal seguente link: [*https://data.worldbank.org/indicator/EN.ATM.CO2E.PC*](https://data.worldbank.org/indicator/EN.ATM.CO2E.PC) ed è composto dalle seguenti variabili:

* *Country Name*: il nome del paese,
* *Country Code*: il codice identificativo del paese,
* *Indicator Name*: l’indicatore che specifica il tipo di gas emesso,
* *Indicator Code*: il codice univoco che rappresenta l’indicatore di riferimento,
* *1960, …, 2023*: 63 variabili che identificano il valore delle emissioni di CO2 misurate in tonnellate metriche per persona per ogni anno considerato,
* *Unnamed: 68*: variabile contenente solo valori nulli.

I due dataset selezionati sono stati fondamentali per comprendere come le emissioni di CO2 variano tra i diversi paesi e per correlare queste informazioni con i dati demografici e territoriali, ottenute precedentemente tramite API, al fine di identificare pattern significativi e relazioni tra popolazione, estensione territoriale e livello di emissioni.

## **Data preparation**

Il processo di Data Preparation include operazioni di pulizia, trasformazione e integrazione dei dati per garantire che siano pronti per l’analisi. In questo progetto, abbiamo suddiviso il processo di Data Preparation in due parti principali: Data Cleansing e Data Integration.

### **2.1 Data cleansing**

Il Data Cleansing consiste nell’identificazione e correzione (o rimozione) dei dati corrotti, inaccurati o irrilevanti.

Le analisi preliminari sui due dataset hanno rilevato che, al contrario del dataset ottenuto tramite API – il quale era già sufficientemente pulito non richiedendo modifiche importanti – il dataset *pollution* necessitava di una pulizia più approfondita. Abbiamo infatti identificato diverse anomalie come valori mancanti e variabili non rilevanti per il nostro studio.

In primo luogo, abbiamo rimosso le colonne che non contenevano informazioni rilevanti per le nostre analisi e le colonne che presentavano unicamente valori nulli. Durante il caricamento dei dati, abbiamo notato infatti che le variabili *Indicator Name* e *Indicator Code* presentavano lo stesso valore ripetuto per tutte le righe del dataset. Inoltre, abbiamo osservato che la variabile *Unnamed: 68* e le variabili riferite agli anni dal 1960 al 1989 e agli anni 2021, 2022 e 2023 presentano valori nulli per tutti i paesi considerati. Per questo motivo abbiamo deciso di eliminare le colonne appena citate ottenendo quindi un dataset contenente unicamente il nome del paese, il codice identificativo del paese e le rilevazioni di CO2 per gli anni dal 1990 al 2020.

Tuttavia, dopo questa pulizia preliminare, abbiamo notato la presenza di ulteriori valori nulli. Più precisamente, 27 paesi non presentavano rilevazioni di emissioni per nessuno degli anni rimanenti e un paese, la Namibia, non presentava invece rilevazioni solamente per il 1990.

È stato quindi deciso di prevedere il valore nullo relativo alla Namibia utilizzando una regressione lineare. Eliminare tutte le osservazioni relative al paese per un unico valore mancante avrebbe infatti causato una perdita di troppe informazioni.

Per quanto riguarda invece i 27 paesi contenenti unicamente valori nulli, abbiamo deciso di mantenere tali valori momentaneamente per rendere più semplice il processo di integrazione dei due dataset. Rimuovere alcuni paesi in questa fase, non avrebbe infatti permesso di individuare correttamente i match nella fase successiva.

### **2.2 Data integration**

La fase di Data Integration descrive in dettaglio il processo di integrazione dei due dataset *pollution* e *data\_from\_mongo* (il DataFrame contenente i dati raccolti tramite API), evidenziando le sfide incontrate e le tecniche impiegate per garantire un abbinamento accurato e coerente.

Per realizzare l'integrazione dei dataset, abbiamo utilizzato il nome del paese come chiave primaria. Considerando la diversità delle fonti, abbiamo riscontrato che non tutti i paesi presenti nel dataset *data\_from\_mongo* erano rappresentati in *pollution* e viceversa, e alcuni nomi non corrispondevano esattamente tra i due dataset. Queste discrepanze hanno richiesto un'attenta valutazione e gestione per garantire che l'integrazione non compromettesse la qualità dei dati.

Per affrontare queste problematiche, abbiamo applicato tecniche di matching, tra cui il fuzzy matching, per identificare e confermare le corrispondenze tra i paesi.

Innanzitutto, tramite la libreria unicode di Python, abbiamo sostituito i caratteri speciali con controparti ASCII nelle variabili dei due dataset contenenti il nome del paese. In questo modo, il processo di matching dovrebbe essere più semplice ed immediato.

Successivamente, utilizzando la libreria fuzzywuzzy di Python, abbiamo comparato i nomi dei paesi e calcolato un punteggio di similarità per ogni possibile abbinamento, permettendo di identificare e correggere discrepanze, come varianti nei nomi o differenze dovute a traduzioni o abbreviazioni. Questo approccio ci ha permesso di identificare un significativo numero di paesi (186 su 250) con corrispondenze dirette, evidenziate da un punteggio di similarità del 100%.

Per i paesi per i quali il fuzzy matching non ha fornito un abbinamento affidabile, abbiamo proceduto con un ulteriore controllo manuale. In alcuni casi, è stato necessario modificare manualmente i nomi dei paesi nei dataset per assicurare che l'integrazione riflettesse le corrispondenze corrette. I paesi per i quali non è stato possibile trovare un match affidabile sono stati esclusi dal dataset finale, decisione guidata dalla necessità di mantenere coerenza e accuratezza nelle analisi e di evitare l'introduzione di incertezze o errori.

Con i nomi dei paesi armonizzati, abbiamo proceduto con l'integrazione fisica dei due dataset utilizzando la funzionalità merge in Pandas. Questa operazione ha creato un unico DataFrame, *combined\_data*, che combinava le informazioni sulle emissioni di CO2 con quelle demografiche e territoriali. La scelta di utilizzare un join interno ha garantito che solo le osservazioni con corrispondenze valide nei due dataset originali fossero inclusi nel dataset finale.

Attraverso questi metodi, abbiamo ottimizzato il processo di integrazione, garantendo che il dataset finale fosse accurato e rappresentativo delle informazioni essenziali per le nostre analisi.

Solo una volta ottenuto il dataset finale pulito e integrato, abbiamo deciso di eliminare i 27 paesi contenenti unicamente valori nulli evidenziati precedentemente per garantire la completezza dei dati per le successive analisi. Il nuovo dataset così ottenuto è stato quindi rinominato con il nome *final\_data.*

## **Data storage**

In seguito alla pulizia e all’integrazione dei due dataset, abbiamo deciso di procedere con la fase di data storage, componente fondamentale per la gestione delle informazioni e cruciale per diversi motivi, prima fra tutti la facilità di accesso, analisi e recupero dei dati.

Prima di caricare il dataset conclusivo *final\_data*, abbiamo svolto alcune operazioni di rifinitura, in modo tale da offrire una lettura dei dati più intuitiva.

Innanzitutto, ci siamo accorti che, dopo l’unione, il dataset finale presentava due colonne per identificare i nomi dei paesi. Abbiamo quindi proceduto con la rimozione dell’attributo *Country Name* in quanto divenuto ridondante.

Inoltre, per semplificare le analisi che ci avrebbero condotto alla risposta delle research questions, abbiamo ritenuto opportuno aggiungere un’ulteriore colonna chiamata *density* rappresentativa della densità di popolazione di ogni paese. Per ottenerla abbiamo diviso i valori contenuti nella variabile *population* per quelli nell’attributo *area*.

Infine, per rendere il dataset più facile da interpretare, abbiamo ordinato gli attributi, mettendo in successione all’inizio quelli contenenti valori descrittivi dei paesi, e in seguito le variabili che indicano le emissioni di CO2 per ogni anno.

La fase di data storage si è quindi conclusa con la trasformazione del dataset finale in dizionario per permettere il caricamento di quest’ultimo, chiamato ora *CO2\_emission*, sul Database Management System NoSQL MongoDB.

Il dataset finale risulta avere i seguenti attributi:

* *\_id*: il codice identificativo dell’osservazione,
* *name*: il nome del paese di riferimento,
* *Country Code*: il codice identificativo del paese di riferimento,
* *population*: la popolazione del paese di riferimento,
* *area*: l’estensione del paese di riferimento in km2,
* *density*: la densità di popolazione paese di riferimento,
* *region*: il continente in cui si trova il paese di riferimento,
* *subregion*: l’area del continente in cui si trova il paese di riferimento,
* *capital*: la capitale del paese di riferimento,
* *languages*: le lingue parlate nel paese di riferimento,
* *borders*: i paesi con cui confina il paese di riferimento,
* *currencies*: la valuta utilizzata nel paese di riferimento,
* *1990, …, 2020*: le emissioni di CO2 in tonnellate metriche pro capite che ogni paese ha emesso dal 1990 al 2020.

## **Data quality**

Prima di procedere con le analisi e prima di rispondere alle domande di ricerca iniziali, è stato di fondamentale importanza verificare che il dataset finale *CO2\_emission* fosse corretto e adeguato.

Innanzitutto, abbiamo deciso di analizzare la **completezza** dei dati per verificare che *CO2\_emission* contenesse tutte le informazioni rilevanti per il nostro studio. Come previsto dalla fase di data preparation, non sono presenti valori nulli in nessuna variabile, ad eccezione della variabile *borders*. Tuttavia, la presenza di valori nulli in questa variabile è ammissibile in quanto non tutti i paesi presentano necessariamente degli stati confinanti, si pensi ad esempio all’Australia che, essendo un’isola, non avrà stati confinanti.

Inoltre, in seguito al processo di integrazione dei dati, è possibile garantire anche la **consistenza** del dataset. I dati raccolti, come è stato analizzato nel paragrafo relativo alla data integration, sono infatti uniformi e coerenti all’interno dello stesso database, nonostante siano stati raccolti da diverse fonti. Non ci sono pertanto discrepanze o contraddizioni tra i dati.

La pulizia, l’eliminazione dei valori nulli o ridondanti e l’integrazione dei dati permette anche di garantire **l’accuratezza** e **l’univocità** dei dati. I dati rispecchiano infatti correttamente e coerentemente la realtà senza presentare duplicati.

La raccolta dei dati da fonti quali la piattaforma Data.Worldbank o dall’API precedentemente mostrato, che continua ad essere periodicamente aggiornato, assicura poi l’**affidabilità** e l’**attualità** dei dati, garantendo dei risultati vicini alla realtà.

Infine, il caricamento del dataset su MongoDB permette un facile accesso e utilizzo all’user. Questo aspetto è fondamentale per poter consentire un’analisi più approfondita dei dati a chiunque ne abbia bisogno.

## **Data analysis**

La fase di data analysis esplora in dettaglio il dataset finale ottenuto per rispondere alle research questions alla base del nostro progetto. Più precisamente, sono stati realizzati due grafici e una tabella che insieme offrono una panoramica significativa delle tendenze globali nelle emissioni di anidride carbonica per capita.

Attraverso il primo grafico, mostrato nella figura 1, abbiamo potuto identificare i 10 paesi con le più alte emissioni di CO2 per capita nel 2020.

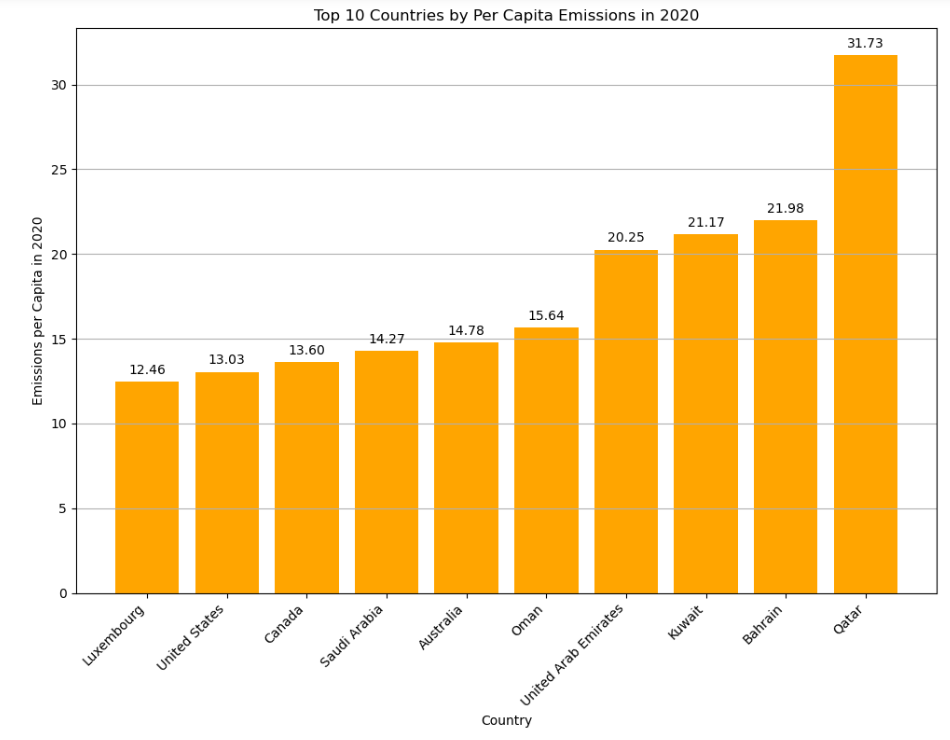


Figura 1 – 10 paesi con le più alte emissioni di CO2 per capita nel 2020

Dal grafico si evince che Qatar, Bahrain e Kuwait registrano le emissioni per capita più elevate. Questo fenomeno potrebbe essere strettamente legato alle loro economie, che sono profondamente radicate nell'uso di risorse energetiche non rinnovabili, come il petrolio e il gas naturale. Questi risultati sottolineano pertanto l'importanza che politiche di sostenibilità ambientale più rigorose e strategie di transizione energetica più precise avrebbero in questi paesi per mitigare l'impatto ambientale.

Parallelamente, la tabella mostrata nella figura 2 mostra la comparazione delle emissioni di CO2 del 1990 e del 2020. I risultati indicano come alcuni paesi abbiano ridotto in modo marcato le loro emissioni nel corso di tre decenni. In particolar modo, il Lussemburgo ha mostrato la riduzione più drastica, passando da 29.55 tonnellate metriche per capita nel 1990 a 12.46 tonnellate nel 2020, segnando una differenza di 17.10 tonnellate. Questo cambiamento potrebbe essere attribuito per esempio a politiche ambientali efficaci e a un incremento nell'uso di energie rinnovabili.

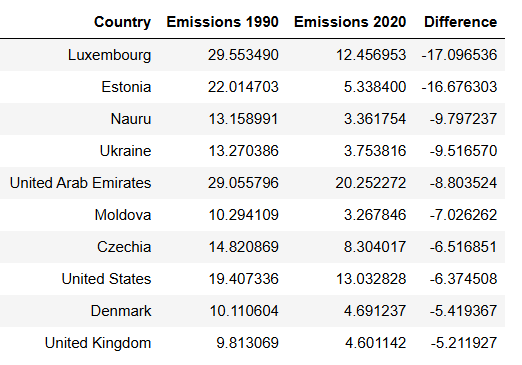


Figura 2 – 10 paesi con la più alta differenza di emissioni di CO2 dal 1990 al 2020

Infine, il secondo grafico (figura 3) ha utilizzato una scala logaritmica per esaminare la relazione tra la popolazione dei paesi e le loro emissioni totali di CO2 nel 2020, offrendo una visione complessiva che sottolinea come la popolazione sia soltanto uno dei fattori che influenzano le emissioni.

Immagine che contiene schermata, diagramma, Diagramma, testo

Descrizione generata automaticamente

Figura 3 – scatter plot tra popolazione e emissioni totali di CO2 nel 2020 per paese

Infatti, mentre paesi asiatici con popolazioni dense, come ad esempio Cina e India, mostrano alte emissioni totali, altri paesi con grandi popolazioni come quelli in Africa non seguono necessariamente questo modello. Questo suggerisce che altri fattori, come potrebbe essere ad esempio l'industrializzazione, svolgono ruoli cruciali nel determinare le emissioni totali.

In conclusione, le analisi condotte forniscono intuizioni significative sulle tendenze globali delle emissioni di CO2 e sottolineano l'importanza di considerare una varietà di fattori quando si valutano le responsabilità ambientali dei paesi. Questi risultati enfatizzano la necessità di strategie differenziate che tengano conto delle specificità regionali ed economiche per combattere efficacemente il cambiamento climatico.

Inoltre, il confronto tra i dati del 1990 e del 2020 mostra come alcune nazioni abbiano fatto progressi significativi nella riduzione delle loro emissioni, mentre altre rimangono altamente inquinanti, delineando un quadro complesso di sfide e progressi nel contesto della sostenibilità globale.

## **Conclusioni**

Il nostro progetto ha dimostrato quanto un efficace processo di data management sia fondamentale per garantire la precisione e l'affidabilità delle analisi ambientali.

Attraverso processi di pulizia, integrazione e analisi dei dati, abbiamo trasformato set di dati grezzi in informazioni precise, pronte per un'interpretazione significativa. Questo ha permesso non solo di ottenere insight accurati sulle emissioni globali di CO2 ma anche di evidenziare come pratiche e politiche specifiche possano influenzarle.

In sintesi, ogni fase del data management è essenziale per supportare decisioni informate e strategie efficaci, confermando che dati ben gestiti sono la chiave per affrontare sfide complesse come il cambiamento climatico.

## **Ruolo dei partecipanti**

**Alessandro Bosi**: identificazione e acquisizione delle fonti di dati necessari per la realizzazione del progetto.

**Denis Bugaenco**: pulizia e integrazione dei dati per creare un dataset uniforme e coerente.

**Eleonora Zullo**: analisi finale dei dati e interpretazione dei risultati.

## **Riferimenti**

**API Rest countries**: https://restcountries.com

**World Bank Data:**  
https://data.worldbank.org/indicator/EN.ATM.CO2E.P

Davy Cielen, Arno D. B. Meysman, and Mohamed Ali. Introducing Data Science, Manning, 2016

Rezzani, *Big Data Analytics*, APOGEO, 2017