Autore: Francesco Marchese

Data: 08/06/2023

Obiettivo: confrontare i dati riguardanti il totale dei positivi, dei ricoveri e dei decessi causati  
 dal Covid-19, in Italia, dal 01/01/2023 sino agli ultimi dati reperibili.

Parte 1 - Punti 1 & 2

Dataset 1

Fonte: <https://github.com/pcm-dpc/COVID-19/tree/master/dati-andamento-nazionale>

Descrizione : repository Github che contiene un importante quantitativo di dati. La raccolta è anche suddivisa in sottocartelle che descrivono l’argomento dei file all’interno mentre ogni file descrive la situazione ad un giorno specifico. Gli ultimi dati sono al 01/06/2023.

Campi : in totale 24, con dei titoli in italiano e descrittivi che agevolano la comprensione dei contagi, i test eseguiti e la loro tipologia

Classificazione : 3 stelle.

------------------------------------------

Dataset 2

Fonte: <https://www.epicentro.iss.it/coronavirus/sars-cov-2-sorveglianza-dati>

Descrizione : file in formato xlsx. Nel complesso è ben fatto, con le schede che indicano il contenuto delle stesse in modo da agevolare la ricerca dei dati interessati. La prima di queste è una nota metodologica con dei link che riportano alle fonti dei dati contenuti nel file. La maggior parte dei dati porta il 20/01/2020 come inizio della raccolta.

Campi : i campi variano in base al foglio che viene consultato, in generale riguardano i casi, i decessi, i ricoveri e l'incidenza in base all'età ed alla condizione clinica. Viene sempre riportata la data di riferimento del dato.

Classificazione : 4 stelle.

------------------------------------------

Dataset 3

Fonte: <https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv>

Descrizione : per ogni stato paese del mondo, viene presentata la situazione giornaliera dal 2020 fino al 2023. Fra tutti i dataset trovati è quello con maggiori casistiche ma con il maggior numero di dati mancanti.

Campi : 67 campi che vanno dai ricoveri, ai casi positivi e decessi fino all'incidenza sulla popolazione per età, condizioni sanitarie etc.

Classificazione : 3 stelle.

------------------------------------------

Dataset 4

Fonte: <https://www.ecdc.europa.eu/en/covid-19/data>

Descrizione : in questo caso ho unito due differenti dataset dalla stessa fonte. I risultato è un dataset che indica per ogni settimana dall'inizio del 2020, tutti i positivi ed i decessi mentre giornalmente i ricoveri in ospedale. Il tutto è calcolato   
 per ogni nazione europea.

Campi : molto basilari e non propriamente ben organizzati. I campi che interessano l'esercizio sono utilizzati come righe, nel caso dei positivi e dei decessi. Gli altri servono ad identificare la nazione e la data.

Classificazione : 3 stelle.

------------------------------------------

Dataset 5

Fonte: <https://www.who.int/data>

Descrizione : per ogni stato, utilizzando la progressione dei giorni, viene descritta la situazione nuovi casi, somma casi, nuovi decessi, somma decessi. Pochi campi ma sono presenti tutti i paesi.

Campi : riguardano l'identificazione della nazione, il giorno del rilevamento, ed il tipo di dato che viene inserito  
Classificazione : 3 stelle.

Tabella nella pagina successiva

Confronto dei dati raccolti dai vari dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Positivi | Ricoveri | Decessi | Data ultima raccolta |
| Dataset 1 | 1198051 | 18402 | 1905171 | 01/06/2023 |
| Dataset 2 | 260846991 | 34688 | 4323 | 01/06/2023 |
| Dataset 3 | 689134 | 18402 | 5600 | 31/05/2023 |
| Dataset 4 | 9538110 | null | 6951 | 28/05/2023 |
| Dataset 5 | 689134 | null | 5474 | 31/05/2023 |

Note

1. Valori cumulativi dall’inizio della raccolta, non è stato possibile definire un periodo.

2. Il valore riguarda il solo giorno 01/06/2023 e non è quindi cumulativo.

Parte 2

Punto 3

Data Engineer: in tutti e 5 i casi si è occupato della raccolta dei dati che, in alcuni casi, è risultato essere veramente gravoso considerando che certi dataset raccolgo informazioni a livello globale.

Data Analyst e Data Scientist: il lavoro più importante è stato fatto sicuramente nel dataset 5 che permette all’utente di esaminare e consultare i dati in maniera grafica ed intuitiva. Negli altri, tolto il lavoro di catalogazione e descrizione della raccolta, non è stato realizzato un importante lavoro grafico.

Punto 4 - I metadati

Nel caso dei Dataset n. 1 e 3, ovvero quelli raccolti in dei repository Github, i metadata non si trovavano internamente il dataset presi in esame ma in file esterni tipo README in cui, in parte, vengono descritti i dati raccolti e come sono somministrati.

Nel caso del Dataset n. 2, la descrizione dei campi è riportata sia sulla pagina web dove è possibile recuperare il dataset che all’interno delle schede del file xlsx.

Nel caso del Dataset n. 4 le descrizioni le descrizioni sono contenute in file separati ma recuperabili sullo stesso sito in cui si trova il dataset.

Nel caso del Dataset n. 5 i metadati sono presenti solo sul sito. I campi sono pochi ma il sito web è ben strutturato ed organizzato e ne rende facile la comprensione.