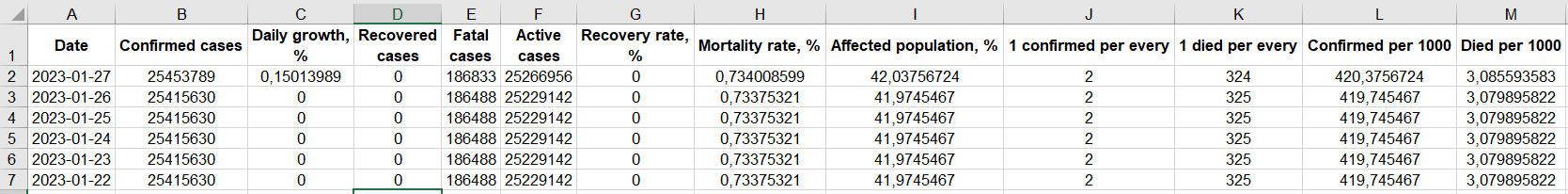
**ESERCITAZIONE DEL 07/02/2023 - DATI COVID**

Per esaminare i dati nazionali di covid19 ho selezionati varie fonti data set, tutti recuperati da internet.

1. Il primo data set è preso da: <https://covid.observer/it/> , sito creato con il focus di fornire tutti i dati covid nel mondo. E’possibile selezionare, però, i dati europei e poi quelli specifici per l’Italia, oltre a poter scaricare i dati grezzi in csv e excel.

Al 27 gennaio 2023 i casi cumulativi da inizio pandemia (2020) risultano : 25.453.789 su una popolazione di 58.870.762 (dato preso da <https://www.populationpyramid.net/it/italia/2023/> ). L’incidenza è quindi di 41,5% quasi 42% .

Il data set riporta le osservazioni dei positivi covid in base alla data **(CHIAVE PRIMARIA**). Gli attributi/colonne sono visibili nell’immagine-esempio sottostante:



In particolare, i dati di interesse per il confronto con la classe sono legati agli attributi: casi positivi, ricoveri, % mortalità.

Valutazione in stelle: il data set in csv è apribile praticamente ovunque ed è possibile modificarlo a proprio piacimento. Non contiene link all’interno del file stesso e come metadati ha solo dimensione e data di creazione nelle proprietà del file, per questo assegnerei 4 stelle. Se si considera, invece, la pagina da cui questi dati sono presi si evidenziano alcune note (anche per spiegare meglio le dashboards) e link alla fonte originaria dei dati, facendo salire il punteggio della pagina a 5 stelle.

1. Il secondo data set è reperibile su internet all’indirizzo: <https://github.com/CSSEGISandData/COVID-19> . E’ una repository github che contiene tutti i i dati relativi al covid forniti dalla Johns-Hopkins University.

I data set sono scaricabili in cvs, excel, json e anche qui la **CHIAVE PRIMARIA** è la data.

I records sono disponibili fino a febbraio e sono cumulativi dal 2020. Al 09/02/2023 i positivi in Italia risultano: 25.488.166 con un’incidenza del 43% che sembra per altro in linea con il dataset precedente. Questo data set contiene alcuni attributi di interesse (paese, numero positivi, ospedalizzati..) ma ha anche molti campi con percentuali e normalizzazioni settimanali.

Valutazione in stelle: 4 stelle se si considera il singolo file. La repository, però, menziona altri link e l’autore dei dati raccolti, se si considera anche questa parte (e non solo il singolo csv) possiamo assegnare 5 stelle.

1. La terza fonte è una testata giornalistica, si tratta di una pagina web del sole 24 ore che ha creato una serie di dashboards riassuntive per monitorare numero di positivi e ospedalizzazioni. Molto interessante la possibilità di scaricare tutti questi dati in csv.

La **CHIAVE PRIMARIA** resta sempre la data mentre gli attributi di interesse sono per lo più gli stessi dei precedenti (positivi totali, ricoveri…) Link alla pagina: <https://lab24.ilsole24ore.com/coronavirus/?refresh_ce>

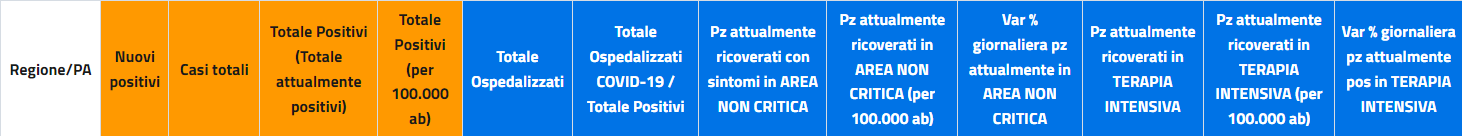
Valutazione in stelle: 3 stelle massimo 4 se si considera il singolo file csv con il metadato relativo alle dimensioni. La pagina, però, completa di dashboard e link vari, può arrivare a 5 stelle

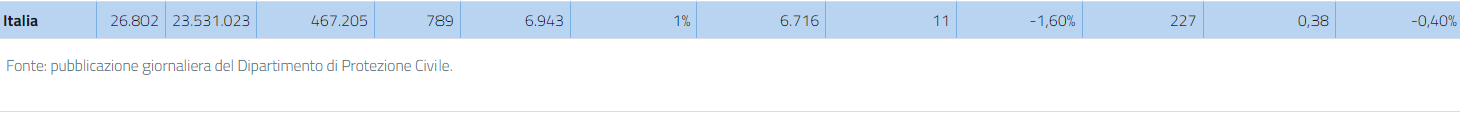
1. La quarta fonte analizzata è la pagina web della fondazione indipendente per la diffusione scientifica: Gimbe <https://coronavirus.gimbe.org/> . I dati riportati purtroppo sono solo visibili in maniera riassuntiva senza possibilità di accedere ai dati grezzi. Resta comunque una buona fonte per effettuare un confronto sommario rispetto i dati totali della classe. **CHIAVE PRIMARIA**: data.

Valutazione in stelle: 1 stella perché non c’è possibilità di scaricare, analizzare, modificare i dati grezzi. Si tratta di una semplice visualizzazione e rielaborazione di dati a cui non abbiamo accesso, bene per fare un confronto con i dati totali della classe ma posso solo visualizzarli.

Un altro esempio ad 1 stella potrebbe essere il semplice report pdf dell’iss al seguente link: <https://www.epicentro.iss.it/coronavirus/bollettino/Bollettino-sorveglianza-integrata-COVID-19_1-febbraio-2023.pdf>

Oppure la tabella della protezione civile al 29/10/2022





1. Un ultimo data set, scaricabile in json, csv, xlsx è reperibile su github: <https://github.com/owid/covid-19-data/tree/02f44ffebe2fc18003efd0a8141a455f38fa7d73/public/data> . i dati riportati sono molti e sono presenti anche record settimali e dati relativi alle vaccinazioni. La **CHIAVE PRIMARIA** è sempre la data.

Valutazione in stelle: 3, massimo 4, se si considera l’unico metadato-proprietà legato alle dimensioni del singolo file csv. Se si considera, invece, come “sistema di controllo” l’intera pagina github allora possiamo notare: menzioni della fonte dei dati, autori, changelog finale per la metologia di lavoro, vari link esterni. In quest’ultimo caso, considerando l’intero sistema si passa a 5 stelle.

Tutti e 5 data set mostrano un’incidenza tra il 42-43%.

**DATI DELLA CLASSE**. I dati sono stati raccolti tramite **sondaggio google** e poi scaricati in csv.

Il file è stato realizzato da Tomasz che ha poi distribuito su discord i dati, qualcuno inoltre li ha poi caricati anche su github.

La **CHIAVE PRIMARIA è l’id** che è stato assegnato numerando in maniera consequenziale le persone partecipanti al sondaggio. Delle 30 persone che hanno compilato il modulo, 18 sono risultate positive ma con 0 ricoveri. Rispetto la media nazionale (circa 42-43%) di casi positivi, **l’incidenza della classe risulta superiore, attestandosi al 60%**. Le ospedalizzazioni, invece, sono nulle per la classe rispetto quelle presenti in Italia. Questo **disallineamento potrebbe dipendere da alcuni fattori:**

1. Età media dei campioni, sicuramente la classe-campione ha un’età ravvicinata e più bassa rispetto l’intera media nazionale.
2. Patologie pregresse. Nei dati nazionali vanno considerate le persone con malattie pregresse (spesso più grandi) che hanno portato ad un aggravarsi della situazione da covid con conseguente ospedalizzazione
3. Dati reali. I dati rilevati a livello nazionali sono quelli ufficiali derivanti da chi ha comunicato di avere il covid o è stato ricoverato per questo motivo, potrebbero però esserci persone che non hanno comunicato l’infezione. Un esempio è dato da chi ha fatto tamponi domiciliari e non avendo gravi sintomi non ha comunicato l’avvenuto contagio.

Il nostro file csv, avendo come metadati l’autore e l’ora di creazione del file (nella stessa pagina csv), può essere classificato a 4 stelle.

Durante la fase di raccolta dei nostri dati abbiamo assunto un ruolo simile a quello del **Data engineer**, definendo i campi di interesse e le modalità di raccolta.

Nell’analizzare i dati e confrontarli con quelli nazionali, il ruolo è stato più simile a quello del **Data Analyst.**

Se avessi fatto una predizione dei contagi futuri o della curva dei contagi avrei invece assunto il ruolo di **Data Scientist**.

Interessante vedere come nel data set riportato dal sole 24 ore (citato al punto 3) si possa intravedere la presenza di un Data Journalist, qualcuno che ha raccolto e raccontato i dati con taglio giornalistico ma scientifico.

In generale, i **metadati** che è stato possibile incontrare durante questa esercitazione erano quelli presenti nelle proprietà del file, es. autore, data creazione, dimensioni. Molto più interessanti però i metadati e i link delle intere pagine da cui sono stati scaricati.