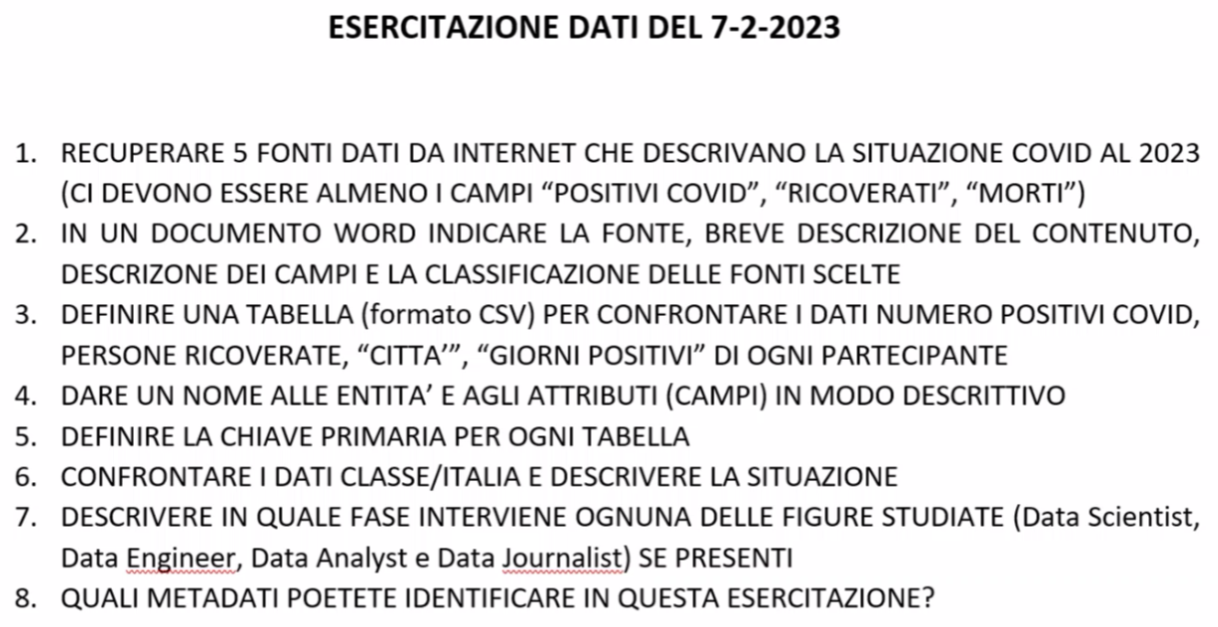
# Esercizio D3 – D4



## Punto 1 - 2:

Elenco fonti che descrivono la situazione Covid:

1. **Fonte:** Istituto superiore di Sanita

<https://www.epicentro.iss.it/coronavirus/sars-cov-2-dashboard>  
a questo linik [COVID-19 ISS open data](https://www.epicentro.iss.it/coronavirus/open-data/covid_19-iss.xlsx) è possibile scaricare un file” xlsx” contenente i dati relativi all’epidemia da Covid-19, aggiornati settimanalmente.   
Pertanto possiamo classificarlo come un file 2 stelle in quanto è un file contenente dati strutturati ma codificati in un formato proprietario (excel appunto). Il vantaggio di questo tipo di dati è che possono essere convertiti facilmente in dati aperti.   
**Campi:** All’interno del file possiamo trovare diversi fogli che descrivono l’andamento della pandemia e un riepilogo del contenuto di ciascun foglio come per esempio i casi per regioni e province, i decessi, i ricoveri, sesso, età, stato clinico ecc.

1. **Fonte:** Our world in Data

<https://ourworldindata.org/coronavirus/country/italy>

Grazie al lavoro di migliaia di ricercatori in questo sito sono riportati dati e informazioni sulle principali tematiche del mondo come fame, malattie, povertà, cambiamento climatico e tra queste anche dati e informazioni sul Covid.   
  
Nella pagina in questione sono presenti due file riassuntivi da scaricare e un link github <https://github.com/owid/covid-19-data> che rimanda diversi dataset.  
  
I due file direttamente scaricabili sono un file xslx e un file csv. Possiamo quindi classificarli rispettivamente con 2 stelle (excel) e 3 stelle (csv) in quanto quest’ultimo contiene dei dati strutturati e codificati in un formato aperto. I file contengono i dati di tutto il mondo ma con un semplice filtro nel file exel è possibile reperire le informazioni relative all’Italia.  
  
**Campi:** iso\_code,continent,location,date,total\_cases,new\_cases,new\_cases\_smoothed,total\_deaths,new\_deaths,new\_deaths\_smoothed,total\_cases\_per\_million,new\_cases\_per\_million,new\_cases\_smoothed\_per\_million,total\_deaths\_per\_million,new\_deaths\_per\_million,new\_deaths\_smoothed\_per\_million,reproduction\_rate,icu\_patients,icu\_patients\_per\_million,hosp\_patients,hosp\_patients\_per\_million,weekly\_icu\_admissions,weekly\_icu\_admissions\_per\_million,weekly\_hosp\_admissions,weekly\_hosp\_admissions\_per\_million,total\_tests,new\_tests,total\_tests\_per\_thousand,new\_tests\_per\_thousand,new\_tests\_smoothed,new\_tests\_smoothed\_per\_thousand,positive\_rate,tests\_per\_case,tests\_units,total\_vaccinations,people\_vaccinated,people\_fully\_vaccinated,total\_boosters,new\_vaccinations,new\_vaccinations\_smoothed,total\_vaccinations\_per\_hundred,people\_vaccinated\_per\_hundred,people\_fully\_vaccinated\_per\_hundred,total\_boosters\_per\_hundred,new\_vaccinations\_smoothed\_per\_million,new\_people\_vaccinated\_smoothed,new\_people\_vaccinated\_smoothed\_per\_hundred,stringency\_index,population\_density,median\_age,aged\_65\_older,aged\_70\_older,gdp\_per\_capita,extreme\_poverty,cardiovasc\_death\_rate,diabetes\_prevalence,female\_smokers,male\_smokers,handwashing\_facilities,hospital\_beds\_per\_thousand,life\_expectancy,human\_development\_index,population,excess\_mortality\_cumulative\_absolute,excess\_mortality\_cumulative,excess\_mortality,excess\_mortality\_cumulative\_per\_million

1. **Fonte:** Presidenza del consiglio dei ministri – Dipartimento della protezione civile

Su Github al seguente link <https://github.com/pcm-dpc/COVID-19> sono presenti numerosi file diverso tipo sull’andamento del Covid di diverso tipo e formato:

* Pdf che possiamo classificare con una stella in quanto anche se disponibili e distribuiti con licenza aperta non sono oggetto di elaborazione, Sono dati leggibili, conservabili e stampabili ma non sono strutturati.
* Csv che classifichiamo con 3 stelle per le motivazioni viste precedentemente
* Json anche in questo caso possiamo classificare l’insieme dei dati con 3 stelle in quanto non sono dotati di URI.
* Xml in quest’ultimo caso invece possiamo attribuire le 4 stelle in quanto all’interno del file vi è la presenza di Uri, in particolare di URL e per alcuni file arriviamo anche alle 5 stelle in quanto i collegamenti presenti riportano ad altri dataset.   
    
  **Campi:** I campi sono più o meno gli stessi trovati negli altri dataset.

1. **Fonte:** Johns Hopkins Whiting School of Engineering

Il Johns Hopkins Coronavirus Resource Center (CRC) è una fonte continuamente aggiornata di dati COVID-19 e indicazioni di esperti che Raccolgono e analizzano i migliori dati disponibili su casi, decessi, test, ricoveri e vaccini per aiutare il pubblico, i responsabili politici e gli operatori sanitari di tutto il mondo a rispondere alla pandemia.   
Al link sul loro sito web (<https://coronavirus.jhu.edu/region/italy>) possiamo vedere subito una presentazione riassuntiva dei dati Covid in Italia. Tutti i dati anche in questo caso sono riportati in diversi file presenti su Github al seguente link <https://github.com/CSSEGISandData/COVID-19> .

È possibile accedere ai file su Github anche dalla Dashboard presente sul sito ArcGIS <https://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6>

I dataset presenti sono tutti in formato csv quindi in questo caso possiamo classificarli con le 3 stelle in quanto non presentano URI. Sono in ogni caso dei dati Open su cui possiamo effettuare delle elaborazioni.

**Campi:** sono presenti tutti i campi elencati fino ad ora.

## Punto 3 – 4 - 5:

Con la classe abbiamo raccolto le nostre informazioni in un file CSV tramite un sondaggio su Google. Questo ci ha permesso di identificare alcune informazioni importanti come il numero di positivi nella classe, il numero di ricoverati e i giorni di degenza dei positivi.   
  
**La chiave primaria** del file è un identificatore numerico progressivo attribuito a ogni partecipante. In questo caso non sarebbe stato per forza necessario un ID in quanto si potevano usare come chiave primaria i campi “nome” e “cognome”. L'importante è che la chiave primaria sia univoca e che identifichi univocamente ogni record all'interno della tabella.  
  
**Gli Attributi** presi in considerazione nel file sono invece i seguenti: ID, Informazioni cronologiche, Nome, Cognome, Città, Positivo, Ricoverato, Giorni Degenza.Punto 6:

Per far un confronto con i dati a livello Nazionale del Covid-19 prenderemo in considerazione il file scaricato dal sito dell’Istituto superiore di Sanita in quanto il più completo a livello di dati.   
  
Innanzitutto è opportuno precisare che il confronto tra i dati sulla pandemia Covid a livello nazionale e un dataset di soli 30 individui non sarebbe statisticamente significativo. I dati sulla pandemia Covid a livello nazionale rappresentano una vasta quantità di informazioni su una popolazione molto più grande, mentre un dataset di solo 30 individui rappresenterebbe solo una piccola frazione di quella popolazione.

Detto questo le informazioni che possiamo evidenziare di questo piccolo Dataset sono le seguenti:

* Il 60% della classe ha contratto il Covid almeno 1 volta.
* Nessuno dei presenti è stato ricoverato.
* I giorni di degenza medi sono stati di 18,67.

Per quanto riguarda i numeri nazionali invece possiamo evidenziare che:

* Il numero dei positivi è di circa 25mil
* Il numero dei ricoveri è di circa 740mila, circa il 3% dei positivi
* Il numero dei deceduti è di circa 180mila e rappresentano circa lo 0,7 dei positivi

## Punto 7:

Considerando le 3 figure principali che lavorano con i dati (Data Engineer, Data Analyst e Data Scientist) potremo avere questo tipo di interventi:

* Data engineer: Il data engineer sarebbe responsabile di raccogliere, pulire e organizzare i dati sulla pandemia Covid. Questo potrebbe includere la creazione di una struttura di database adeguata per i dati, la trasformazione dei dati in un formato utilizzabile e la gestione dei flussi di dati da diversi fonti.
* Data analyst: Il data analyst sarebbe responsabile di esplorare i dati e di analizzare le tendenze e le relazioni presenti all'interno di essi. Questo potrebbe includere la creazione di visualizzazioni dei dati, l'utilizzo di statistiche descrittive per identificare le tendenze e la creazione di modelli per prevedere l'evoluzione della pandemia.
* Data scientist: Il data scientist sarebbe responsabile di utilizzare metodi avanzati di analisi dei dati, come l'apprendimento automatico, per prevedere l'evoluzione della pandemia e identificare fattori che influiscono sulla sua diffusione. Potrebbe anche utilizzare tecniche di modellizzazione per valutare le politiche pubbliche e le strategie di prevenzione e controllo della malattia.

In generale, questi tre ruoli lavorano insieme per ottenere una comprensione più profonda dei dati e per utilizzare questa conoscenza per prendere decisioni informate su come gestire la pandemia.  
  
Punto 8:  
  
I metadati inclusi in questi file potrebbero essere diversi come:

* Nome del file
* Data di creazione
* Fonte dei dati: la fonte originale dei dati, ad esempio il Ministero della Salute italiano o un'altra agenzia governativa
* Intervallo di tempo dei dati: il periodo di tempo coperto dai dati nel file
* Definizione dei campi: la descrizione dei campi presenti nel file, ad esempio il numero di casi, il numero di decessi, il numero di persone vaccinate, ecc.
* Definizione delle unità di misura: le unità di misura utilizzate per descrivere i dati, ad esempio il numero di casi per 100.000 abitanti
* Definizione delle variabili: la descrizione delle variabili presenti nel file, ad esempio la regione, la provincia, la data, etc.
* Metodo di raccolta dei dati: la descrizione del metodo utilizzato per raccogliere i dati, ad esempio la raccolta manuale o la raccolta automatizzata da fonti online
* Limiti dei dati: la descrizione di eventuali limiti dei dati, ad esempio la mancanza di dati sui casi asintomatici o la mancata copertura di alcune aree geografiche
* Note: eventuali note sulle fonti dei dati, sulla loro accuratezza o sulla loro attendibilità
* Le autorizzazioni d’accesso