Lecture 3 - esercizio

COVID

1. RECUPERARE 5 FONTI DATI DA INTERNET CHE DESCRIVANO LA SITUAZIONE COVID AL 2023 (CI DEVONO ESSERE ALMENO I CAMPI "POSITIVI COVID", "RICOVERATI", "MORTI")
2. IN UN DOCUMENTO TESTO INDICARE LA FONTE, BREVE DESCRIZIONE DEL CONTENUTO, DESCRIZIONE DEI CAMPI E LA CLASSIFICAZIONE 5 STELLE DELLE FONTI SCELTE
3. DEFINIRE UNA TABELLA PER CONFRONTARE I DATI "NUMERO POSITIVI" COVID, "PERSONE RICOVERATE", "CITTA'", "GIORNI POSITIVI" DI OGNI PARTECIPANTE
4. DARE UN NOME ALLE ENTITA' E AGLI ATTRIBUTI (CAMPI) IN MODO DESCRITTIVO

COVID – PARTE 2

1.
    1. ~~DEFINIRE LA CHIAVE PRIMARIA PER OGNI TABELLA~~
    2. CONFRONTARE I DATI CLASSE/ITALIA E DESCRIVERE LA SITUAZIONE
    3. DESCRIVERE IN QUALE FASE INTERVIENE OGNUNA DELLE FIGURE STUDIATE (Data Scientist, Data Engineer, Data Analyst e Data Journalist) SE PRESENTI
    4. QUALI METADATI POTETE IDENTIFICARE IN QUESTA ESERCITAZIONE?

Parte 1:

Prima fonte:
https://ourworldindata.org/explorers/coronavirus-data-explorer?zoomToSelection=true&time=2020-03-01..latest&facet=none&country=~ITA&pickerSort=asc&pickerMetric=location&Metric=Confirmed+cases&Interval=7-day+rolling+average&Relative+to+Population=true&Color+by+test+positivity=false
IL contenuto del dataset e' interattivo e di facile accesso:
Il dataset puo' essere fltrato per nazione, intervallo di tempo, si possono utlizzare svariate metriche per visualizzare il dataset (le piu' importanti: morti confermate, casi confermati, casi in ospedalieri, casi in terapia intensiva, test, varie metriche sui vaccini ecc.) e si possono scegliere vari metodi di visualizzazione (se con un istogramma o in forma mappale).
Il dataset puo' essere scaricato in un file csv, e i metadati ci sono.
Cinque stelle : i dati sono strutturati, sono accessibili attraverso URL e sono collegati ad altri database LOD

Seconda fonte:
https://www.ecdc.europa.eu/en/covid-19/data
Questo e' piu un insieme di dataset che un singolo dataset, poiche' questo ente ha deciso di fare un lavoro piu' diramato, infatti i dati per casi e morti sono in dataset separato rispetto a quelli per i dati ospedalieri (ricorevati, guariti). Questo rispetto al dataset precedente non e' interattivo ma allo stesso modo e' totalmente accessibile e con metadati che aiutano nella navigazione del dataset.
Come il dataset precedente il dataset riguarda dati nazionali.
Cinque stelle : i dati sono strutturati, sono accessibili attraverso URL e sono collegati ad altri database LOD

Terza fonte:
https://health.google.com/covid-19/open-data/explorer?loc=IT
Questa come la prima fonte e' in una fonte che ha un tipo di visualizzazione interattiva in forma tabellare ed e' la prima fonte di quelle elencate che presente dati per le singole citta' Italiane ma questi ultimi sono aggiornati fino al 15/09/2022.
I campi nella versione interattiva sono: periodo di analisi (ultimi 7 gg, 30 gg ecc), la nazione da cui poi si puo selezionare la regione e citta' e il relativo numero di casi e decessi, mentre sono disponibili anche vaccinazioni, guariti e ricoveri, ma solo per nazione.
Sono presenti i metadati e il file csv per esplorare il dataset.
Nel file .csv del dataset troviamo una quantita' di dati molto piu' ampia come il numero di infermieri su 1000 persone, l'umidita' media, il numero di scuole chiuse, la social mobility (che non ho ben capito cosa sia ma e' un dato che veniva aggiornato).
Cinque stelle : i dati sono strutturati, sono accessibili attraverso URL e sono collegati ad altri database LOD

Quarta fonte:
https://covid19.who.int/data
https://app.powerbi.com/view?r=eyJrIjoiYWRiZWVkNWUtNmM0Ni00MDAwLTljYWMtN2EwNTM3YjQzYmRmIiwidCI6ImY2MTBjMGI3LWJkMjQtNGIzOS04MTBiLTNkYzI4MGFmYjU5MCIsImMiOjh9
Questo e' il dataset piu' schematico che ho trovato, sul download del dataset c'e' una piccola tabella che specifica esattamente che tipo di dato si trova nella tabella se STRING o INT o DATE e una relativa descrizione, anche gli altri dataset presentavano queste caratteristiche ma non in una maniera cosi diretta, dovevi un minimo andare a cercare il metadato.

Ci sono dati come nuovi casi, nuove morti, casi cumulativi, morti cumulativi, nazione. In questo dataset non sono presenti pero' purtroppo i dati sui ricoveri e guariti
Cinque stelle : i dati sono strutturati, sono accessibili attraverso URL e sono collegati ad altri database LOD

Quinta fonte:
https://dati-covid.italia.it:
Questo e' il dataset migliore per lo scopo di fare una analisi a livello nazionale poiche' ci sono dati molto aggiornati e molto dettagliati a livello territoriali ma anche a livello di orario (sono aggiornati anche ogni ora) pero' rispetto agli altri e' anche il piu' difficile da lavorare con gli strumenti che sono in mio posseso in questo momento, sia a livello tecnico che a livello di conoscenza poich'e ci sono file .csv per ogni giorno, quindi per fare un analisi di un anno mi servirebbe aprire 365 file.
Ci sono a livello nazionale dati su ricoverati su sintomi, ospedalizzati, positivi, dimessi, variazione totale positivi ecc.
A livello regionale ci sono totali casi, dimessi, ospedalizzati, deceduti ed altri.
I metadati sono presenti sottoforma di file .xml ma non so interpretarli.
quattro stelle poiche' mancano comllegamenti a databese esterni LOD

https://app.powerbi.com/view?r=eyJrIjoiYWRiZWVkNWUtNmM0Ni00MDAwLTljYWMtN2EwNTM3YjQzYmRmIiwidCI6ImY2MTBjMGI3LWJkMjQtNGIzOS04MTBiLTNkYzI4MGFmYjU5MCIsImMiOjh9

Svolgimento tabella:
Ora rilevazione dati 09/06/2023 17.00
La tabella non e' assolutamente delle migliori, ma penso rispecchi un minimo anche la mia poca esperienza nello svilupparle.
La navigazione dei miei dati non e' molto intuitiva, per un motivo principale, per un discorso di completezza ho deciso di fare un analisi nel 2022, ma allo stesso tempo i dati erano parziali su tre fonti, poiche' per health_google e covid_italia, le rilevazioni si concludevano rispettivamente il 17/09/2022 e il 25/11/2022.
Dato che quindi ho diciamo per tutti la stessa data di inizio rilevazione (01/01/2022) mentre ho tre diverse date di fine rilevazione (31/12/2022 - 17/09/2022 - 25/11/2022) ci sono quattro colonne relative alle date, mentre le voci casi_periodo_ricoveri e morti_periodo_ricoveri si riferiscono ai casi in Italia nel periodo di rilevazione dei ricoveri (es casi_periodo_ricoveri di google health si riferisce al numero dei casi in Italia rilevati dal 01/01/2022 al 17/09/2022)
I dati si distanziano l'uno dall'altro in un maniera che posso definire (nella mia esperienza) non significativa considerando i vari metodi di raccolta dati tranne in numero di casi di ourworld in data che e' visibelmente un outlier visto che e' piu' grande di un ordine di 10 volte rispetto agli altri e non sono riuscito a trovare motivazioni sul perche'.

Ho notato che nella formattazione in .pdf i valori dell'ultima colonna sono visibili, ma non il nome del campo, per evitare di fare un .txt preferisco annotare qui che l'ultimo valore e' ricoveri_25_11_2022 cioe fondamentalmente la data di fine rilevazione di covid italia

	casi_periodo_ricoveri	morti_periodo_ricoveri	ricoveri_01_01_2022	ricoveri_31_12_2022	ricoveri_17_09_2022	ricover
ourworldindata	190427550	475450	85850	39160	NULL	NULL
ecdc.europa	17644956	49312	8868	4200	NULL	NULL
health_google	11786731	137402	12410	NULL	3775	NULL
WHO	19067148	47459	11231	3887	NULL	NULL
covid italia	17832171	43166	12562	7350	NULL	7350

Parte 2:

Prendero' questo dataset per spiegare le tre figure https://covid19.who.int/data

- Data Engineer: Nel caso del dataset del WHO, e' riempito da un sistema chiamato The Oxford Covid-19 Government Response Tracker (OxCGRT) che colleziona informazioni dai governi mondiali e quando e li trasferisce nel dataset, questo sistema di collezione e trasferimento dei dati e' fatto dal Data Engineer.
- Data Analyst: Ha preso le informazioni e le ha lavorate in maniera grafica nel sito https://ourworldindata.org (si questo e' un po una confessione perche' la mia prima fonte ha utilizzato la mia quarta fonte per fare parte del suo lavoro)
- Data Scientist: Analizzava i dati per tutto il periodo e faceva stime sull'avvenire dell'andamento delle casistiche.

Da quello che ho capito di cosa e' un metadato, quasi tutto l'esercizio e' un metadato perche' danno tutte informazione di navigazione dei dataset.