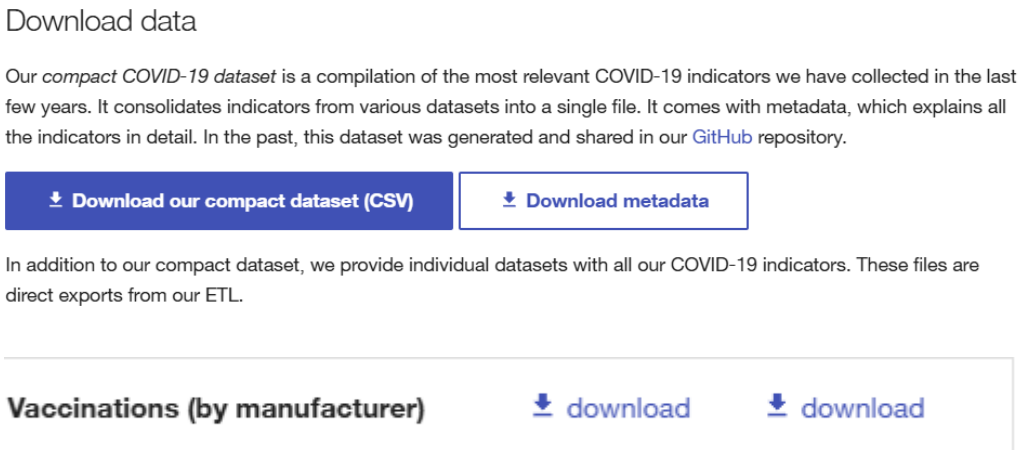
PROJECT OVERVIEW, LINKS AND COLUMNS

1 Overview and Links to Databases of COVID-19 Project

This project offers a comprehensive analytical exploration of the COVID-19 pandemic using two key public data sources: **Our World in Data (OWID)** and the **Vaccine Adverse Event Reporting System (VAERS)**. The main goal is to combine large-scale global data with detailed individual-level reports of adverse events to generate insightful dashboards and visual narratives in Power BI. You can download both databases from the attached links. I'm also adding screenshots to show the buttons you have to press to download them:

A) OWID compact.csv and vaccinations_manufacturer.csv:



B) VAERS data.csv , symptoms.csv and vax.csv:

Year	Zip File	CSV File (VAERS DATA)	CSV File (VAERS Symptoms)	CSV File (VAERS Vaccine)
All Years Data*	543.25 MB			
2025*	4.48 MB	16.85 MB	1.54 MB	1.57 MB
2024	13.65 MB	50.78 MB	5.20 MB	5.13 MB
2023	25.75 MB	102.32 MB	11.39 MB	10.14 MB
2022	64.80 MB	274.73 MB	27.84 MB	21.94 MB
2021	175.80 MB	647.83 MB	81.48 MB	60.03 MB
2020	11.78 MB	43.94 MB	4.82 MB	4.70 MB

The ETL process was designed from scratch, using **Python (pandas)** for cleaning and transformation, and **Power BI** for modeling and visualization. All CSVs were cleaned, transformed, and in some cases automated via custom Python functions for scalability. The project includes over **20 CSV files across 5**

different formats, covering country-level pandemic trends, vaccine administration by manufacturer, and detailed clinical reports of adverse events post-vaccination.

There are **two core star models**:

- One centered on **OWID_Compact** (global pandemic facts) connected to a custom calendar dimension.
- Another centered on **VAERS_Data** (US vaccine adverse events) linked via VAERS_ID to related vaccine and symptom tables.

This work is not about drawing causal conclusions on vaccine safety. Instead, it is focused on **tracking real-world reports** and allowing comparative epidemiological analyses over time, geography, demographics, vaccine types, and symptom clusters.

Data preparation was not merely technical—it involved **critical thinking**. Many countries were excluded due to poor reporting, especially in parts of Africa and Asia, which can lead to significant **underestimation** of real **pandemic impact**. Filtering and normalization were necessary to ensure **analytical robustness** and fair comparisons.

All datasets were prepared with **Power BI modeling in mind** (e.g., category creation, ranking, calculated columns and over **100 DAX measures**), and they enable:

- **Time-series analysis** (smoothed metrics, YTD and LY trends)
- **Severity and outcome distribution** (deaths, hospitalizations, disabilities)
- **Ranking and performance indicators** by country, region, and vaccine type
- **Efficiency scores** based on multiple normalized metrics
- **Custom unpivot transformations** to analyze symptom frequencies

The final deliverables include interactive **dashboards**, **automated** data pipelines, and extensive **documentation** (both technical and analytical).

2 Description and Cleaning of Columns

2.1 OWID Dataset compact.csv

Total columns: 60
Columns removed: 23
Columns kept: 37

A) 🌍 Geographical and Temporal Information

Column	Description	Keep	Reason
country	Country name	✅	Essential for grouping and comparisons
code	Country code (ISO Alpha-3)	❌	Redundant unless standardization is needed
continent	Continent to which the country belongs	✅	Enables regional aggregation
date	Date of the data	✅	Core time variable for trends

B) □ COVID-19 Cases

Column	Description	Keep	Reason
total_cases	Cumulative confirmed cases	✅	Shows the total evolution of the pandemic
new_cases	Daily new cases	❌	Too noisy; replaced with smoothed version
new_cases_smoothed	7-day rolling average of new cases	✅	Shows real trends, filters daily reporting fluctuations
total_cases_per_million	Total cases per million people	✅	Allows comparisons between countries regardless of population size
new_cases_per_million	Daily new cases per million	❌	Redundant; replaced by smoothed version per million
new_cases_smoothed_per_million	Smoothed new cases per million	✅	Ideal for normalized trend comparisons

C) ☠COVID-19 Deaths

Column	Description	Keep	Reason
total_deaths	Cumulative deaths	✅	Key indicator of pandemic impact
new_deaths	Daily new deaths	✅	Useful for daily monitoring
new_deaths_smoothed	Smoothed new deaths	❌	Redundant if using both daily and per million metrics

total_deaths_per_million	Deaths per million	✅	Enables normalized cross-country comparison
new_deaths_per_million	Daily deaths per million	✅	Adjusted by population
new_deaths_smoothed_per_million	Smoothed deaths per million	❌	Redundant with existing metrics

D) 🏷️ Excess Mortality

Column	Description	Keep	Reason
excess_mortality	Weekly % of excess mortality	✅	Indicates excess deaths beyond expectations
excess_mortality_cumulative	Cumulative % of excess mortality	✅	Shows prolonged deviation over time
excess_mortality_cumulative_absolute	Absolute number of excess deaths	✅	Gives real volume of excess mortality
excess_mortality_cumulative_per_million	Excess deaths per million	✅	Allows cross-national analysis

E) 🏥 Hospitalizations and ICU

Column	Description	Keep	Reason
hosp_patients	Current hospitalized patients	❌	Redundant; prefer normalized version
hosp_patients_per_million	Hospitalized per million people	✅	Enables fair comparison across countries
weekly_hosp_admissions	Weekly hospital admissions (raw)	❌	Not normalized; better to use per million
weekly_hosp_admissions_per_million	Weekly hospital admissions per million	✅	Reflects burden over time
icu_patients	Current ICU patients	❌	Redundant with ICU admissions

icu_patients_per_million	ICU patients per million	✗	Less useful if ICU admissions are used
weekly_icu_admissions	Weekly ICU admissions (raw)	✗	Same reason as above
weekly_icu_admissions_per_million	ICU admissions per million	✅	Captures severe case trends

F) □ Testing and Positivity

Column	Description	Keep	Reason
total_tests	Total tests performed	✗	Hard to interpret without positives
new_tests	Daily new tests	✗	Too volatile
total_tests_per_thousand	Total tests per 1000 people	✅	Shows overall testing effort
new_tests_per_thousand	New tests per 1000 people	✗	Redundant; replaced by smoothed version
new_tests_smoothed	Smoothed new tests	✅	Tracks real testing trends
new_tests_smoothed_per_thousand	Smoothed new tests per 1000 people	✅	Enables normalized comparison
positive_rate	Percentage of positive tests	✅	Key to assessing under-testing
tests_per_case	Number of tests per case detected	✅	Inversely related to positivity rate

G) 💉 Vaccination

Column	Description	Keep	Reason
total_vaccinations	Total doses administered	✗	Redundant if focusing on individuals
people_vaccinated	People with at least one dose	✗	Dropped for simplification
people_fully_vaccinated	Fully vaccinated people	✗	Dropped for simplification

total_boosters	Total booster doses	✅	Important for booster analysis
new_vaccinations	Daily new doses	❌	Replaced by smoothed version
new_vaccinations_smoothed	Smoothed new doses	✅	Shows true vaccination pace
total_vaccinations_per_hundred	Doses per 100 people	❌	Redundant with individual-based metrics
people_vaccinated_per_hundred	One dose per 100 people	❌	Redundant
people_fully_vaccinated_per_hundred	Fully vaccinated per 100 people	✅	Key indicator of national coverage
total_boosters_per_hundred	Boosters per 100 people	❌	Redundant
new_vaccinations_smoothed_per_million	Smoothed new doses per million	❌	Dropped for simplification
new_people_vaccinated_smoothed	Smoothed new individuals vaccinated	❌	Dropped by design
new_people_vaccinated_smoothed_per_hundred	Per 100 people	❌	Dropped by design

H) □ Demographic and Health Indicators

Column	Description	Keep	Reason
population	Total population	✅	Needed for all per capita calculations
population_density	Population density	✅	Helps understand transmission dynamics
median_age	Median age	✅	Important for assessing mortality risk
life_expectancy	Life expectancy	✅	General health indicator

gdp_per_capita	GDP per capita (USD)	✅	Socioeconomic indicator
extreme_poverty	% living in extreme poverty	✅	Crucial for assessing access to healthcare
diabetes_prevalence	Diabetes prevalence (%)	✅	Known comorbidity for COVID-19
handwashing_facilities	% with access to handwashing	✅	Indicator of basic hygiene and prevention capacity
hospital_beds_per_thousand	Hospital beds per 1000 people	✅	Measures healthcare capacity
human_development_index	Human Development Index (HDI)	✅	Overall development and wellbeing metric

2.2 OWID Dataset vaccinations_manufacturer.csv

Total columns: 4
Columns removed: 0
Columns kept: 4

Column	Description	Keep	Reason
country	Country name	✅	Essential for grouping
vaccine	Vaccine name	✅	Key for comparison between vaccines
date	Date of administration	✅	Required for timeline and trends
total_vaccinations	Cumulative doses administered	✅	Core quantitative indicator of usage

2.3 VAERS Dataset data.csv

Total columns: 35
Columns removed: 18
Columns kept: 17

A) □ Identification and Dates

Column	Description	Keep	Reason
VAERS_ID	Unique report identifier	✅	Primary key to link across VAERS files

RECVDATE	Report received date	✅	Useful to analyze reporting trends
RPT_DATE	Report creation date	❌	Often empty or redundant with RECVDATE
TODAYS_DATE	File creation date (not event date)	❌	Irrelevant for analytical purposes
VAX_DATE	Date of vaccination	✅	Needed to calculate time to symptom onset
ONSET_DATE	Symptom onset date	✅	Crucial to assess reaction time
NUMDAYS	Days between vaccination and symptoms	✅	Already calculated; saves preprocessing steps
DATEDIED	Date of death (if applicable)	✅	Required for severity and fatality timeline analysis

B)Demographics

Column	Description	Keep	Reason
STATE	US state where the event occurred	✅	Useful for regional analysis
AGE_YRS	Patient age in years	✅	Key for age-based segmentation
CAGE_YR	Age in text format (years)	❌	Redundant with numeric age (AGE_YRS)
CAGE_MO	Age in months (for infants)	❌	Not needed if analysis excludes infants or already using AGE_YRS
SEX	Patient gender (M/F/U)	✅	Fundamental for demographic breakdowns

C) Clinical Outcome (Severity)

Column	Description	Keep	Reason
DIED	Indicates death	✅	Key indicator of event severity
L_THREAT	Life-threatening condition	✅	Useful for classification of serious cases
ER_VISIT	Emergency room visit	✅	Acts as proxy for acute impact
ER_ED_VISIT	Emergency visit (duplicate)	❌	Redundant with ER_VISIT
HOSPITAL	Was hospitalized	✅	Helps measure clinical burden
HOSPDAYS	Days spent in hospital	✅	Useful for estimating impact severity
DISABLE	Resulted in disability	✅	Captures long-term adverse effects

RECOVD	Patient recovered	✅	Allows tracking of recovery vs. severity
BIRTH_DEFECT	Birth defect caused	✅	Rare but critical for completeness

D) 🤒Symptom Description

Column	Description	Keep	Reason
SYMPTOM_TEXT	Free text description of symptoms	❌	Unstructured and redundant with codified symptom tables

E) Patient Medical Context

Column	Description	Keep	Reason
LAB_DATA	Lab data (free text)	❌	Unstructured; difficult to scale
V_ADMINBY	Who administered the vaccine (e.g., pharmacy)	❌	Operational, not analytically useful
V_FUNDBY	Funding source (e.g., public/private)	❌	Contextual info, not needed for clinical analysis
OTHER_MEDS	Current medication (free text)	❌	Not standardized
CUR_ILL	Current illnesses (free text)	❌	Difficult to interpret systematically
HISTORY	Medical history (free text)	❌	Lacks structure for consistent analysis
PRIOR_VAX	Previous vaccines received	❌	Inconsistent and loosely related to COVID vaccine
SPLTTYPE	Report type (e.g., VAERS internal)	❌	Administrative metadata, not useful for analysis
FORM_VERS	Form version	❌	Technical detail with no analytical value

2.4 VAERS Dataset symptoms.csv

Total columns: 11
Columns removed: 5
Columns kept: 6

Column	Description	Keep	Reason

VAERS_ID	Unique report identifier (shared across VAERS)	✅	Key for joining with VAERS Data and VAX tables
SYMPTOM1	First reported symptom (usually the most severe)	✅	Allows analysis of primary adverse reactions
SYMPTOMVERSION1	Dictionary version for symptom 1	❌	Internal technical detail, not analytically relevant
SYMPTOM2	Second reported symptom	✅	Expands scope of adverse event analysis
SYMPTOMVERSION2	Dictionary version for symptom 2	❌	Redundant and technical
SYMPTOM3	Third reported symptom	✅	Helps analyze co-occurring symptoms
SYMPTOMVERSION3	Dictionary version for symptom 3	❌	No analytical value
SYMPTOM4	Fourth reported symptom	✅	Adds clinical context
SYMPTOMVERSION4	Dictionary version for symptom 4	❌	Repetitive, not useful
SYMPTOM5	Fifth reported symptom	✅	Covers up to 5 symptoms per report
SYMPTOMVERSION5	Dictionary version for symptom 5	❌	Purely technical, no impact on analysis

2.5 VAERS Dataset vax.csv

Total columns: 9
Columns removed: 5
Columns kept: 4

Column	Description	Keep	Reason
VAERS_ID	Unique report ID	✅	Key to merge with SYMPTOMS and DATA tables
VAX_TYPE	Vaccine type (e.g., COVID19, FLU)	✅	Required to filter only COVID-related reports
VAX_MANU	Manufacturer (Pfizer, Moderna, etc.)	✅	Important for brand-level comparison of adverse effects
VAX_LOT	Vaccine lot number	❌	Highly specific, often missing, not suitable for broad analysis
VAX_DOSE_SERIES	Dose number (1st, 2nd, booster)	✅	Enables analysis by dose sequence
VAX_ROUTE	Route of administration (e.g., IM)	❌	Too technical, not useful for impact evaluation

VAX_SITE	Injection site (e.g., left/right arm)	✕	Does not influence severity of reaction
VAX_NAME	Full commercial vaccine name	✕	Redundant with VAX_TYPE and VAX_MANU

3 Glossary of Key Terms and Columns

Term	Description
OWID_Compact	Main OWID dataset with daily country-level metrics (cases, deaths, tests, vaccinations, etc.)
VAERS_Data	Patient-level US data on post-vaccination adverse events (age, gender, hospitalization, death, etc.)
VAERS_Symptoms	Table listing up to 5 symptoms per reported adverse event (unpivoted in Power BI)
VAERS_VAX	Information about vaccine manufacturer, dose number, and type
new_cases_smoothed	7-day rolling average of new COVID-19 cases
new_deaths_per_million	Daily deaths per million people
total_boosters	Total booster doses administered per country
positive_rate	Proportion of tests that are positive (key for testing adequacy)
tests_per_case	Number of tests performed for each confirmed case
Days from Vaccination to Death	Calculated column: difference between vaccination date and date of death
VAX_DOSE_SERIES	Number indicating whether the vaccine was the 1st, 2nd, or a booster dose
DISABLE, HOSPITAL, DIED	Binary columns (Y/N) indicating adverse event severity
Daily Doses	Calculated daily vaccine doses per manufacturer (non-cumulative)
Log Vaccinations	Logarithmic transformation of total vaccinations (to improve visual scaling)
Stringency Index	OWID metric representing government intervention strength
Pandemic Efficiency Score	Composite metric created from positivity rate, vaccination, testing, and deaths
Dim_Calendar	Custom date table for connecting all datasets in Power BI
Rank by...	Ranking metrics used to compare countries by various health or policy metrics
People Fully Vaccinated per Hundred	Percentage of population with full vaccination protocol