

Exploratory Data Analysis (EDA) of COVID-19 Deaths, Infections, and Vaccinations in Global Perspective

Overview

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has brought the world to a standstill and emerged as one of the most significant global health crises in recent history. Since its emergence in late 2019, the virus has rapidly spread across continents, affecting millions of people and disrupting economies, healthcare systems, and daily life. Its unprecedented impact has resulted in a staggering loss of life and profound socio-economic consequences. Governments, healthcare professionals, and researchers worldwide have been working tirelessly to understand the virus, develop effective treatments and vaccines, and implement measures to control its transmission.

The following report aims to extract valuable insights using SQL queries and provide a comprehensive analysis of the Covid-19 dataset. Concurrently, to explore the global impact of the pandemic by examining the recorded fatalities, infections, and vaccine quantities across all nations.

Key Metrics and Statistics

Source: <https://ourworldindata.org/covid-deaths>

Date: 01-Jan-2020 – 14-May-2023

Total number of records analyzed: 309,667

Global Population: 8,045,247,353

Techniques

Various SQL queries ranging from simple to complex were employed to perform EDA on the dataset. A combination of aggregations, filtering, grouping, joining, and CTEs were used to extract insights from the data.

Queries and Findings

1. Countries with no deaths

- A query was made to find out countries that suffered the least due to COVID-19 with the use of aggregations and grouping.

```
SELECT LOCATION, SUM(NEW_DEATHS) DEATH_COUNT
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL
GROUP BY LOCATION
HAVING SUM(NEW_DEATHS) = 0
```

Results:

	LOCATION	DEATH_COUNT
1	Saint Helena	0
2	Tuvalu	0
3	Niue	0
4	Pitcairn	0
5	Tokelau	0
6	North Korea	0
7	Turkmenistan	0
8	Vatican	0
9	Falkland Islands	0

2. First reported case and death for each country

- A query was made to identify the first case and first death reported in each country. In this case, analyzing results for India with the use of aggregations and grouping

```
SELECT CONTINENT, LOCATION, MIN(NEW_CASES) CASES
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL AND NEW_CASES <> 0 and location = 'India'
GROUP BY CONTINENT, LOCATION
ORDER BY 3 ASC

SELECT CONTINENT, LOCATION, MIN(NEW_DEATHS) DEATHS
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL AND NEW_DEATHS <> 0 and location = 'India'
GROUP BY CONTINENT, LOCATION
ORDER BY 3 DESC
```

Results:

	CONTINENT	LOCATION	DATE	CASES
1	Asia	India	2020-01-30 00:00:00.000	1

	CONTINENT	LOCATION	DATE	DEATHS
1	Asia	India	2020-03-13 00:00:00.000	1

3. Countries with highest number of covid-19 deaths

- The query below helps identify which country was affected the most in terms of death with the use of aggregations and grouping

```
SELECT CONTINENT, LOCATION, MAX(TOTAL_DEATHS) TOTAL_DEATH_COUNT
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL
GROUP BY CONTINENT, LOCATION
HAVING MAX(TOTAL_DEATHS) IS NOT NULL
ORDER BY 3 DESC
```

Results: (viewing first 10 results only)

	continent	LOCATION	total_death_count
1	North America	United States	1125209
2	South America	Brazil	701833
3	Asia	India	531707
4	Europe	Russia	398578
5	North America	Mexico	333960
6	Europe	United Kingdom	225081
7	South America	Peru	220196
8	Europe	Italy	189904
9	Europe	Germany	173473
10	Europe	France	163120

4. Case fatality rate (CFR)

- The following query assists in determining the count of individuals who contracted the virus and subsequently lost their lives with the use of aggregations and grouping

```
SELECT LOCATION, POPULATION, SUM(NEW_CASES) TOTAL_CASES, SUM(NEW_DEATHS) TOTAL_DEATHS,
(MAX(TOTAL_DEATHS)/MAX(TOTAL_CASES))*100 AS CASE_FATALITY_RATE
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL
GROUP BY LOCATION, POPULATION
HAVING SUM(NEW_CASES) IS NOT NULL AND SUM(NEW_CASES)<>0
ORDER BY 5 DESC
```

Results: (viewing first 10 results only)

	LOCATION	POPULATION	TOTAL_CASES	TOTAL_DEATHS	CASE_FATALITY_RATE
1	Yemen	33696612	11945	2159	18.0745081624111
2	Sudan	46874200	63993	5046	7.88523744784586
3	Syria	22125242	57423	3163	5.5082458248437
4	Somalia	17597508	27334	1361	4.97914685007683
5	Peru	34049588	4503222	220196	4.88974338817851
6	Egypt	110990096	516023	24830	4.81180102437294
7	Mexico	127504120	7595575	333960	4.39677106693977
8	Bosnia an...	3233530	402890	16339	4.0554493782422
9	Afghanistan	41128772	217361	7902	3.63542677849292
10	Liberia	5302690	8091	295	3.63411619283066

5. Mortality rate

- The following query checks the rate of deaths per million in each country with the use of aggregations and grouping

```
SELECT LOCATION, POPULATION, (SUM(NEW_DEATHS)/POPULATION)*1000000 MORTALITY_RATE_PER_MILLION
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL
GROUP BY LOCATION, POPULATION
HAVING (SUM(NEW_DEATHS)/POPULATION)*1000000 IS NOT NULL
ORDER BY 3 DESC
```

Results: (viewing first 10 results only)

	LOCATION	POPULATION	MORTALITY_RATE_PER_MILLION
1	Peru	34049588	6466.92112691643
2	Bulgaria	6781955	5653.38460665103
3	Bosnia and Herzegovina	3233530	5052.99162215906
4	Hungary	9967304	4892.49650657791
5	North Macedonia	2093606	4621.69099630016
6	Georgia	3744385	4555.8883501563
7	Croatia	4030361	4518.95003946297
8	Montenegro	627082	4477.88327523354
9	Slovenia	2119843	4395.13680965996
10	Czechia	10493990	4076.80967868275

6. Infection rate

- The following query checks the number of people per million infected with the virus in each country with the use of aggregations and grouping

```
SELECT LOCATION, (SUM(NEW_CASES)/MAX(POPULATION))*1000000 AS INFECTION_RATE_PER_MILLION
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL
GROUP BY LOCATION
HAVING (SUM(NEW_CASES)/MAX(POPULATION))*1000000 IS NOT NULL
ORDER BY 2 ASC
```

Results: (viewing first 10 results only)

	LOCATION	INFECTION_RATE_PER_MILLION
1	Turkmenistan	0
2	North Korea	0
3	Yemen	354.486676583391
4	Niger	362.981018530919
5	Chad	434.343197253425
6	Tanzania	657.701962045965
7	Sierra Leone	901.9579180041
8	Burkina Faso	972.754237011552
9	Democratic Republic of Congo	973.66720218043
10	Nigeria	1220.25037144481

7. Comparing Infection Rate (IR) with case fatality rate (CFR)

- This query compares the percentage of the population infected with the virus against the percentage of people who were infected that lost their lives with the use of aggregations and grouping

```
SELECT LOCATION, POPULATION, (SUM(NEW_CASES)/MAX(POPULATION))*100 AS INFECTION_RATE,
(MAX(TOTAL_DEATHS)/MAX(TOTAL_CASES))*100 AS CASE_FATALITY_RATE
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL
GROUP BY LOCATION, POPULATION
HAVING MAX(TOTAL_CASES) <> 0
ORDER BY 4 ASC
```

Results: (viewing first 10 results only)

	LOCATION	POPULATION	INFECTION_RATE	CASE_FATALITY_RATE
1	Falkland Islands	3801	50.5919494869771	0
2	Tokelau	1893	0.264131008980454	0
3	Vatican	808	3.21782178217822	0
4	Niue	1952	38.2684426229508	0
5	Pitcairn	47	8.51063829787234	0
6	Tuvalu	11335	24.5169827966476	0
7	Saint Helena	5401	40.1036845028698	0
8	Nauru	12691	42.4946812701915	0.0185425551641016
9	Burundi	12889583	0.416925822968827	0.0279121697059918
10	Cook Islands	17032	41.4924847346172	0.0283005518607613

8. Countries with infection rate greater than 50%

- This query checks for the total number of countries where more than 50% of their population were infected with the virus with the use of CTEs, aggregations and grouping

```
WITH CTE AS(
    SELECT LOCATION, MAX(TOTAL_CASES) TOTAL_CASES, MAX(POPULATION) POPULATION, (MAX(TOTAL_CASES)/MAX(POPULATION))*100 INFECTION_RATE
    FROM COVIDDEATHS
    WHERE CONTINENT IS NOT NULL
    GROUP BY LOCATION
)

SELECT COUNT(LOCATION) LOCATION
FROM CTE
WHERE INFECTION_RATE>=50
```

Results:

	LOCATION
1	25

9. Death counts per continent

- The query identifies the continent most impacted in terms of loss of lives with the use of aggregations and grouping

```
SELECT CONTINENT, MAX(POPULATION) POPULATION, SUM(NEW_DEATHS) TOTAL_DEATH_COUNT
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL
GROUP BY CONTINENT
ORDER BY 3 ASC
```

Results:

	CONTINENT	POPULATION	TOTAL_DEATH_COUNT
1	Oceania	26177410	26534
2	Africa	218541216	258922
3	South America	215313504	1355057
4	North America	338289856	1602085
5	Asia	1425887360	1631628
6	Europe	144713312	2059668

10. Spread Pattern

- This query gives us an overview of how the virus spread globally by providing the location of when every case was reported since it started

```
SELECT *
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL AND NEW_CASES>0
ORDER BY 4 ASC, 3 ASC
```

Results: (viewing first 10 results only)

	iso_code	continent	location	date	population	total_cases	new_cases	total_deaths	new_deaths
1	CHN	Asia	China	2020-01-04 00:00:00.000	1425887360	1	1	0	0
2	FIN	Europe	Finland	2020-01-04 00:00:00.000	5540745	1	1	0	0
3	DEU	Europe	Germany	2020-01-04 00:00:00.000	83369840	1	1	0	0
4	CHN	Asia	China	2020-01-06 00:00:00.000	1425887360	4	3	0	0
5	MCO	Europe	Monaco	2020-01-08 00:00:00.000	36491	1	1	0	0
6	FIN	Europe	Finland	2020-01-09 00:00:00.000	5540745	2	1	0	0
7	ESP	Europe	Spain	2020-01-11 00:00:00.000	47558632	1	1	0	0
8	CHN	Asia	China	2020-01-12 00:00:00.000	1425887360	45	41	1	1
9	ESP	Europe	Spain	2020-01-12 00:00:00.000	47558632	2	1	0	0
10	THA	Asia	Thailand	2020-01-13 00:00:00.000	71697024	1	1	0	0

11. Global Overview

- This query analyzes the global cases and death count along with infection and death rate with the use of aggregations.

```
SELECT SUM(NEW_CASES) TOTAL_CASES, SUM(NEW_DEATHS) TOTAL_DEATHS,
(SUM(NEW_CASES)/MAX(POPULATION))*100 GLOBAL_INFECTION_RATE,
(SUM(NEW_DEATHS)/SUM(NEW_CASES))*100 GLOBAL_DEATH_RATE
FROM COVIDDEATHS
WHERE CONTINENT IS NOT NULL
```

Results:

	TOTAL_CASES	TOTAL_DEATHS	GLOBAL_INFECTION_RATE	GLOBAL_DEATH_RATE
1	765919550	6933894	53.7152913677557	0.905303174465255

12. Total vaccinations per Country

- The below query identifies the total vaccinations received by the population of each country with the use of aggregations and grouping

```
SELECT CONTINENT, LOCATION, MAX(TOTAL_VACCINATIONS) TOTAL_VACCINATIONS
FROM COVIDVACCINATIONS
WHERE CONTINENT IS NOT NULL
GROUP BY CONTINENT, LOCATION
ORDER BY 3 DESC
```

Results: (viewing first 10 results only)

	CONTINENT	LOCATION	TOTAL_VACCINATIONS
1	Asia	China	3491077000
2	Asia	India	2206662991
3	North America	United States	676728782
4	South America	Brazil	486436436
5	Asia	Indonesia	444303130
6	Asia	Japan	383747738
7	Asia	Bangladesh	358040699
8	Asia	Pakistan	338642674
9	Asia	Vietnam	266266588
10	North America	Mexico	223158993

13. Total vaccinations per Continent

- The below query identifies the total vaccinations received around each continent with the use of aggregations and grouping

```
SELECT CONTINENT, MAX(TOTAL_VACCINATIONS) TOTAL_VACCINATIONS
FROM COVIDVACCINATIONS
WHERE CONTINENT IS NOT NULL
GROUP BY CONTINENT
ORDER BY 2 ASC
```

Results:

	CONTINENT	TOTAL_VACCINATIONS
1	Oceania	63681652
2	Africa	116606863
3	Europe	192221468
4	South America	486436436
5	North America	676728782
6	Asia	3491077000

14. Impact of vaccinations on death rates

- The below query analyzes how increase in vaccinations across India impacted its death rates with the use of inner joins

```
SELECT CD.LOCATION, CD.DATE, CD.NEW_DEATHS, CV.TOTAL_VACCINATIONS
FROM COVIDDEATHS CD
JOIN COVIDVACCINATIONS CV
ON CD.LOCATION = CV.LOCATION AND CD.DATE = CV.DATE
WHERE CD.CONTINENT IS NOT NULL AND CD.LOCATION = 'INDIA'
ORDER BY 2 desc
```

Results: (viewing first 10 results only)

	LOCATION	DATE	NEW_DEATHS	TOTAL_VACCINATIONS
1	India	2023-05-13 00:00:00.000	NULL	2206662991
2	India	2023-05-12 00:00:00.000	NULL	2206661076
3	India	2023-05-11 00:00:00.000	NULL	2206659444
4	India	2023-05-10 00:00:00.000	NULL	2206657460
5	India	2023-05-09 00:00:00.000	15	2206655406
6	India	2023-05-08 00:00:00.000	12	2206652521
7	India	2023-05-07 00:00:00.000	21	2206651926
8	India	2023-05-06 00:00:00.000	17	2206649602
9	India	2023-05-05 00:00:00.000	36	2206648123
10	India	2023-05-04 00:00:00.000	22	2206646421

15. Death rates vs vaccinations

- This query analyzes the death rates in each country and compares them to the total vaccinations with the use of inner joins

```
SELECT CV.LOCATION, MAX(CD.TOTAL_DEATHS) TOTAL_DEATHS, MAX(CV.PEOPLE_VACCINATED) TOTAL_VACCINATIONS
FROM COVIDVACCINATIONS CV
JOIN COVIDDEATHS CD
ON CV.LOCATION = CD.LOCATION AND CD.DATE = CV.DATE
WHERE CV.CONTINENT IS NOT NULL
GROUP BY CV.LOCATION
ORDER BY 2 DESC
```

Results: (viewing first 10 results only)

	LOCATION	TOTAL_DEATHS	TOTAL_VACCINATIONS
1	United States	1125209	270227181
2	Brazil	701833	189643431
3	India	531707	1027406053
4	Russia	398578	88798804
5	Mexico	333960	97179493
6	United King...	225081	53813491
7	Peru	220196	30484622
8	Italy	189904	50895645
9	Germany	173473	64876299
10	France	163120	54673331

16. Vaccinations overview

- This query gives an overview of the vaccinations received globally with the use of inner joins and CTEs

```
WITH CTE AS(
SELECT POPULATION, SUM(CD.NEW_DEATHS) DEATHS , MAX(CV.PEOPLE_VACCINATED) VACCINES
FROM COVIDVACCINATIONS CV
JOIN COVIDDEATHS CD
ON CV.LOCATION = CD.LOCATION AND CD.DATE = CV.DATE
WHERE CV.CONTINENT IS NOT NULL
GROUP BY POPULATION
)

SELECT SUM(POPULATION) GLOBAL_POPULATION, SUM(DEATHS) GLOABL_DEATH_COUNT, SUM(VACCINES) GLOBAL_VACCINATION_COUNT,
(SUM(VACCINES)/SUM(POPULATION)) * 100 GLOBAL_VACCINATION_RATE
FROM CTE
```

Results:

	GLOBAL_POPULATION	GLOABL_DEATH_COUNT	GLOBAL_VACCINATION_COUNT	GLOBAL_VACCINATION_RATE
1	8045247353	6933894	5635438968	70.0468080188808

17. Yearly COVID-19 Data

- The below query compares the data yearly for India with the use of inner joins

```
SELECT CD.LOCATION LOCATION, YEAR(CD.DATE) AS YEAR, SUM(CD.NEW_DEATHS) DEATHS, MAX(CV.PEOPLE_VACCINATED) VACCINES
FROM COVIDDEATHS CD
JOIN COVIDVACCINATIONS CV
ON CD.LOCATION = CV.LOCATION AND CD.DATE = CV.DATE
WHERE CD.CONTINENT IS NOT NULL AND CD.LOCATION ='INDIA'
GROUP BY CD.LOCATION, YEAR(CD.DATE)
ORDER BY 1 ASC, 2 ASC
```

Results:

	LOCATION	YEAR	DEATHS	VACCINES
1	India	2020	148738	NULL
2	India	2021	332342	845640601
3	India	2022	49622	1027200953
4	India	2023	1005	1027406053

Insights

Covid-19 emerged as a global pandemic that impacted and reshaped our world in unprecedented ways. It affected many countries globally and took the lives of many. Among the countries impacted the most, United States emerged as the country with the highest death count of 1,125,209 followed by Brazil with 701833 deaths and India with 531707 deaths.

At a proper glance, continent wise, the 10 countries with the highest deaths were mostly located in the European continent. Despite the severity of the virus, about 9 countries emerged as being the least impacted as these countries reported 0 death cases.

It is also necessary to examine how the virus initially started to spread across the globe. On 04/01/2020, the first ever COVID-19 case was reported in China. On the same day, Finland and Germany reported a case. Just 2 days after, China reported another case. It can be seen that the spread of the virus was contained within Asia and Europe until the United States reported its first case on the 20/01/2020.

It is notable that Tokelau, Saint Helena and Pitcairn were among the last countries to report a case. Tokelau held the distinction of being the last nation to confirm a case while the Marshall Islands were the last to report a Covid-19 death.

When analyzing India alone, the country with the third highest death count and second largest population in the world, this country reported its first death on the 13/03/2020 just 2 months after it had reported its first COVID-19 case on the 30/01/2020. This provides an insight into the rapidity with which the virus distributed within countries.

The case fatality rate (CFR) gives us an impression of how many people who contracted the virus, had died due to it. Yemen, unfortunately, had a CFR of about 18%, the highest of all countries.

The mortality rate (MR), on the other hand, calculates the number of deaths per million. Upon querying, it was found that 6466 people per million were dying in Peru. With its population of over 34 million, it can be approximated that the number of deaths in Peru were over 200,000. Despite having the highest mortality rate, it ranked as the 7th country with the highest death count.

Furthermore, the infection rate gives us an idea of the number of people per million that were infected with the virus. North Korea and Turkmenistan had an infection rate of 0 while Cyprus had the highest. As a result, about 737554 individuals in every million were infected with the virus.

Upon comparing the infection rates with the case fatality rates of every country, it can be seen that about 7 countries whose population had been infected with the virus managed to have a case fatality rate of 0. At the same time, there were about 25 countries where more than half of its population had been infected with the virus.

On a wider scale, when analyzing the death count across each continent, Oceania reported the least number of deaths whereas Asia reported the highest.

Collectively, the virus had a substantial global influence, leaving an enduring mark. To date, the cumulative count of reported cases stands at a staggering 760 million, indicative of an infection rate nearing 53 percent. Furthermore, the global tally of lives lost to the virus approximates 7 million, underscoring the profound gravity of its consequences.

Given the vast global impact of the virus, it became imperative for nations to pursue a preventative measure against the virus. Consequently, as a means to effectively counteract the virus, vaccination initiatives were introduced.

Vaccination efforts were extensively deployed across all nations, with China leading as the recipient of the largest number of doses, approximating 3.4 billion. Following closely were India and the United States. Notably,

the Asian region witnessed the highest quantity of vaccinations administered while the Oceania region reported the lowest.

To assess the effectiveness of these vaccination campaigns on large countries such as India, it is essential to evaluate their impact on mortality rates. The initiation of India's vaccination campaign occurred on January 16, 2020. Prior to this milestone, the mortality rates were quite elevated. This could potentially reflect infections that developed prior to the vaccination. Notably, as the vaccination coverage expanded, the mortality rates took a significant downward trend. The success of the vaccination program can be further recognized by evaluating the yearly death count. From 2020 to early 2021, which was prior to the initiation of the vaccination program, the death counts had increased by 12%. However, between 2021 to 2023, the death counts dropped remarkably by 99%. This substantial decline, displays the significant success of the vaccination campaign as a pivotal preventative measure against the virus.

To grasp the extensiveness at which the vaccination program took place, it is noteworthy that out of the total global population of 8,045,247,353, about 5,635,438,968 individuals had received vaccine administrations. This signifies that roughly 70% of the world's population took part in the collective effort to fight against the virus.

Conclusion

The global impact of the COVID-19 pandemic has been significant and far-reaching with numerous countries experiencing high death counts and widespread infections. The data presented highlights both the devastating loss of lives and the unbelievable imprint left across the globe. Vaccination efforts emerged as a critical turning point, significantly reducing mortality rates and marking a huge impact on the population of every country. Understanding the severity of the pandemic's impact globally, there is still a need for countries to continue to prioritize preventive measures such as vaccination campaigns, testing, and strengthening of healthcare infrastructure. Moreover, sharing best practices and collaborating internationally still remain crucial as it yields a more effective global response to the ongoing pandemic and future public health crises.