The Covid Project

March, 19th, 2023

By Craig Schlachter

Step 1: Ask

In this step, we define the problems and objectives of our case study and its desired outcome.

1.1 Background

COVID-19 is a disease that has affected many people in most places. But, the severity of its impact has been different between countries, with some countries effectively limiting infections and deaths.

This project’s aim is to conduct a descriptive analysis of the global and country-level response to identify countries that performed well and those that didn't. This analysis will provide key statistics and summarize the response, serving as a platform for further study on this topic.

1.2 Business Task

Analyze a comprehensive COVID-19 dataset to gain insights into the global management of the coronavirus, which countries responded well and the others who didn't.

1.3 Business Objectives

- What are the global infections, vaccinations, deaths, and death percentage?
- What is the death count per continent?
- What percentage of each country has been infected, vaccinated, and likelihood of death?
- What are the trends identified?

1.4 Deliverables

- A clear summary of the business task
- A description of all data sources used
- Documentation of any cleaning or manipulation of data
- A summary of analysis
- Supporting visualizations and findings
- High-level recommendations based on the analysis

1.5 Key Stakeholders

- The key stakeholders of this project are all the individuals who might be interested in understanding the impact of the pandemic on their local communities or making informed decisions about their personal health and safety.

. . .

Step 2: Prepare

In the Prepare phase, we identify the data being used and its limitations.

2.1 Information on Data Source

1. Data is publicly available on Our World in Data: COVID-19 Data. Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, Esteban Ortiz-Ospina and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]
2. All raw data is sourced and collected from the World Health Organization from January 2020 - March 2023.
3. All significant countries, territories and areas of Earth have submitted COVID-19 data.
4. Data collected includes location, date period, new cases, new vaccinations, new deaths and countless more significant data points.

2.2 Limitations of Data Set

- Caution must be taken when interpreting all data presented, and differences between information products published by WHO, OWD, national public health authorities, and other sources using different inclusion criteria and different data cut-off times are to be expected.

2.3 Is Data ROCCC?

A good data source is ROCCC which stands for Reliable, Original, Comprehensive, Current, and Cited.

- Reliable - High - Reliable as it includes all countries, territories, and areas
- Original - High - Data is sourced from World Health Organization
- Comprehensive - High - Parameters cover wide range of variables
- Current - High - Updated daily
- Cited - High - One of the world's leading public health journals.

Overall, this dataset is considered reliable, high quality data and it is endorsed for producing recommendations based on this data.

2.4 Data Selection

The following file is downloaded and then imported into our created SQL tables 'covid_deaths' and 'covid_vaccinations'.

2.5 Tools

We are using SQL for exploratory data analysis. Finally, we are using Tableau for visualizations.

Step 3: Process

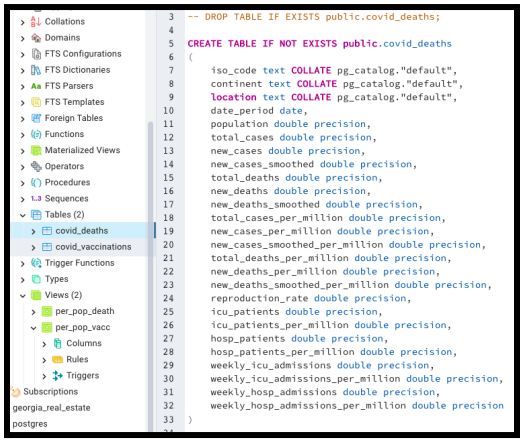
Here, we will process the data to ensure it is clean, correct, relevant, complete and free of errors and outliers by performing:

- Explore and observe the data
- Check for and treat missing or null values
- Transform data - format data type
- Perform preliminary statistical analysis

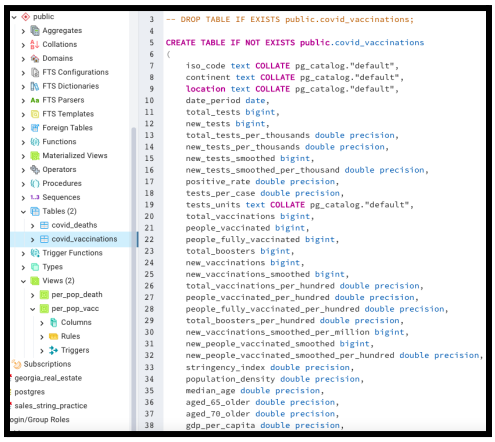
3.1 Preparing the environment

The SQL tables are created, columns are named, and data types are set.

[covid_deaths]

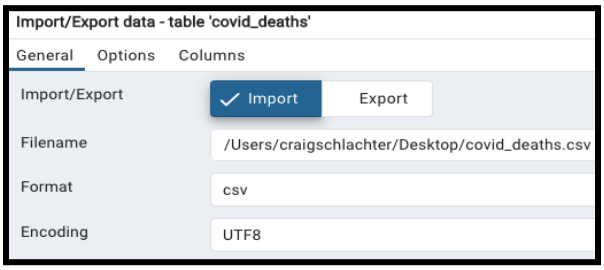


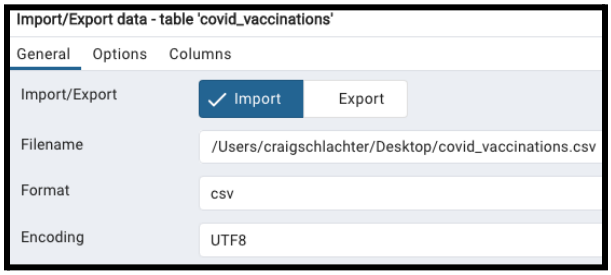
[covid_vaccinations]



3.2 Importing Dataset

Reading in the selected file.

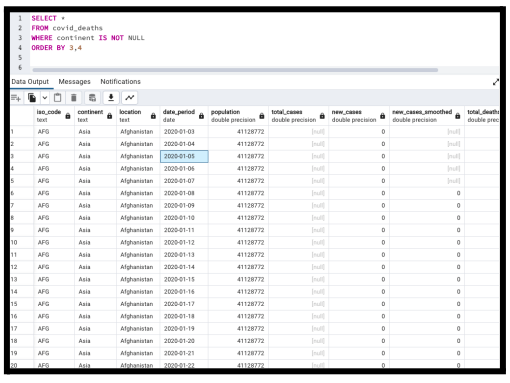


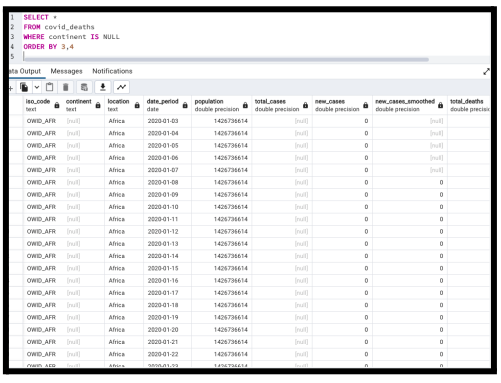


3.2 Data Cleaning and Manipulation

1. Observe and familiarize with data
2. Check for nulls or missing values
3. Perform validation checks of data

Previewing the dataset to get familiarized with the data.





While looking at the data we noticed a significant characteristic of our dataset. When the **continent column is null**(table on the right), it populates the location column with the continent and its values. Going forward we'll filter out the null continents so we don't get erroneous calculations for individual countries.

. . .

Step 4: Analyze

4.1 Performing Calculations

Pulling statistics for analysis:

1. Selecting data we are going to be using.
2. Calculating total cases vs. total deaths to see the likelihood of dying if you contract covid in your country.
3. Quantifying total cases vs. total population to look at the percentage of the population that got covid.
4. Calculating the countries with the highest infection rate compared to population.
5. Computing the countries with the highest death count.
6. Determining the continents with the highest death count.
7. Quantifying the daily global new cases, new deaths, and death percentage.
8. Computing the current global total cases, total deaths, and death percentage.
9. Calculating total vaccinations vs. total population to look at the percentage of the population vaccinated
10. Determine the rolling count of people vaccinated and percentage of people vaccinated on a per day basis.
11. Evaluate the rolling count of people dead and percentage of people dead on a per day basis.

****Going forward, I will show the results for the above calculations for the United States for illustrative purposes only unless stated otherwise.*

[Step 1's Results]

```sql
SELECT location,
       date_period,
       total_cases,
       new_cases,
       total_deaths,
       population
FROM covid_deaths
WHERE continent IS NOT NULL AND location = 'United States'
ORDER BY 1,2
```

Data Output | Messages | Notifications

	location text	date_period date	total_cases double precision	new_cases double precision	total_deaths double precision	population double precision
1	United States	2020-01-03	[null]	0	[null]	338289856
2	United States	2020-01-04	[null]	0	[null]	338289856
3	United States	2020-01-05	[null]	0	[null]	338289856
4	United States	2020-01-06	[null]	0	[null]	338289856
5	United States	2020-01-07	[null]	0	[null]	338289856
6	United States	2020-01-08	[null]	0	[null]	338289856
7	United States	2020-01-09	[null]	0	[null]	338289856
8	United States	2020-01-10	[null]	0	[null]	338289856
9	United States	2020-01-11	[null]	0	[null]	338289856
10	United States	2020-01-12	[null]	0	[null]	338289856
11	United States	2020-01-13	[null]	0	[null]	338289856
12	United States	2020-01-14	[null]	0	[null]	338289856
13	United States	2020-01-15	[null]	0	[null]	338289856
14	United States	2020-01-16	[null]	0	[null]	338289856
15	United States	2020-01-17	[null]	0	[null]	338289856
16	United States	2020-01-18	[null]	0	[null]	338289856
17	United States	2020-01-19	[null]	0	[null]	338289856

[Step 2's Results]

```sql
-- Looking at Total Cases vs Total Deaths
-- Shows the likelihood of dying if you contract covid in your country
SELECT location,
       date_period,
       total_cases,
       total_deaths,
       (total_deaths/total_cases)*100 AS death_percentage
FROM covid_deaths
WHERE location = 'United States'
ORDER BY 1,2
```

Data Output | Messages | Notifications

	location text	date_period date	total_cases double precision	total_deaths double precision	death_percentage double precision
56	United States	2020-02-27	56	[null]	[null]
57	United States	2020-02-28	64	[null]	[null]
58	United States	2020-02-29	69	1	1.4492753623188406
59	United States	2020-03-01	73	1	1.36986301369863
60	United States	2020-03-02	82	2	2.4390243902439024
61	United States	2020-03-03	100	2	2
62	United States	2020-03-04	135	8	5.9259259259259265
63	United States	2020-03-05	186	11	5.913978494623656
64	United States	2020-03-06	256	13	5.078125
65	United States	2020-03-07	334	16	4.790419161676647
66	United States	2020-03-08	464	19	4.094827586206897
67	United States	2020-03-09	610	24	3.934426229508197
68	United States	2020-03-10	822	28	3.40632603406326
69	United States	2020-03-11	1212	32	2.6402640264026402
70	United States	2020-03-12	1709	39	2.2820362785254535
71	United States	2020-03-13	2234	51	2.2829006266786034
72	United States	2020-03-14	2961	58	1.9587977034785546

[Step 3's Results]

```sql
-- Looking at Total Cases vs population
-- Shows what percentage of population got covid
SELECT location,
       date_period,
       population,
       total_cases,
       (total_cases/population)*100 AS pop_covid_percentage
FROM covid_deaths
WHERE location = 'United States'
ORDER BY 1,2
```

Data Output | Messages | Notifications

	location text	date_period date	population double precision	total_cases double precision	pop_covid_percentage double precision
228	United States	2020-08-17	338289856	5414874	1.600661061501058
229	United States	2020-08-18	338289856	5456369	1.6129271697700565
230	United States	2020-08-19	338289856	5497958	1.6252210648610168
231	United States	2020-08-20	338289856	5539583	1.637525601713579
232	United States	2020-08-21	338289856	5586157	1.6512930851819572
233	United States	2020-08-22	338289856	5632344	1.6649461697131114
234	United States	2020-08-23	338289856	5679525	1.6788930851062824
235	United States	2020-08-24	338289856	5722075	1.6914710561111237
236	United States	2020-08-25	338289856	5758285	1.7021748946560193
237	United States	2020-08-26	338289856	5794001	1.7127327045833738
238	United States	2020-08-27	338289856	5834575	1.724726561118049
239	United States	2020-08-28	338289856	5881005	1.7384514775400182
240	United States	2020-08-29	338289856	5927925	1.752321240161573
241	United States	2020-08-30	338289856	5974471	1.7660804467042606
242	United States	2020-08-31	338289856	6015379	1.7781730351382454
243	United States	2020-09-01	338289856	6052014	1.789002505590945
244	United States	2020-09-02	338289856	6090710	1.8004121168953716

[Step 4's Results(Comprehensive List)]

```sql
-- Looking at countries with highest infection rate compared to population

SELECT location,
       population,
       MAX(total_cases) AS highest_infection_count,
       MAX((total_cases)/population)*100 AS pop_covid_percentage
FROM covid_deaths
WHERE continent IS NOT NULL AND total_cases IS NOT NULL
GROUP BY 1,2
ORDER BY pop_covid_percentage DESC
```

Data Output | Messages | Notifications

	location text	population double precision	highest_infection_count double precision	pop_covid_percentage double precision
1	Cyprus	896007	650685	72.62052640213749
2	San Marino	33690	23616	70.0979519145147
3	Austria	8939617	5943417	66.48402274952048
4	Faeroe Islands	53117	34658	65.24841387879586
5	Slovenia	2119843	1330654	62.77134674596184
6	Gibraltar	32677	20433	62.53022003243872
7	Martinique	367512	229020	62.31633252791745
8	Brunei	449002	279661	62.28502322929519
9	Andorra	79843	47890	59.98021116441016
10	Jersey	110796	66391	59.92183833351385
11	South Korea	51815808	30581499	59.01963161512409
12	Saint Pierre and Miquelon	5885	3415	58.028887000849615
13	Denmark	5882259	3403958	57.86821015531618
14	France	67813000	38538948	56.83120935513839
15	Greece	10384972	5896788	56.781934510752 65
16	Iceland	372903	209137	56.08348551768154
17	Guernsey	63329	34991	55.252727818219135

[Step 5's Results(Comprehensive List)]

```sql
-- Showing countries with highest death count

SELECT location,
       MAX(total_deaths) AS total_death_count
FROM covid_deaths
WHERE continent IS NOT NULL AND total_deaths IS NOT NULL
GROUP BY 1
ORDER BY total_death_count DESC
```

Data Output | Messages | Notifications

	location text	total_death_count double precision
1	United States	1111342
2	Brazil	699276
3	India	530775
4	Russia	396378
5	Mexico	333100
6	Peru	219493
7	United Kingdom	207695
8	Italy	188322
9	Germany	168583
10	France	161397
11	Indonesia	160934
12	Iran	144902
13	Colombia	142629
14	Argentina	130472
15	China	120227
16	Spain	119479
17	Poland	118970

[Step 6's Results(Comprehensive List)]

```sql
-- We take these out as they are not inluded in the above queries and want to stay consistent
-- European Union is part of Europe

SELECT location,
      SUM(new_deaths) AS total_death_count
FROM covid_deaths
WHERE continent IS NULL
AND location NOT IN ('World','European Union','International','High income','Upper middle income','Lower
AND new_deaths IS NOT NULL
GROUP BY 1
ORDER BY 2 DESC
```

Data Output | Messages | Notifications

	location text	total_death_count double precision
1	Europe	2023354
2	Asia	1624789
3	North America	1595874
4	South America	1351340
5	Africa	258757
6	Oceania	25338

[Step 7's Results(Global List)]

```sql
-- Global Numbers

SELECT date_period,
       SUM(new_cases) AS total_new_cases,
       SUM(new_deaths) AS total_new_deaths,
       SUM(new_deaths)/SUM(new_cases)*100 AS death_percentage
FROM covid_deaths
--WHERE location = 'United States'
WHERE continent IS NOT NULL AND new_cases <> 0
GROUP BY 1
ORDER BY 1
```

Data Output | Messages | Notifications

	date_period date	total_new_cases double precision	total_new_deaths double precision	death_percentage double precision
45	2020-02-23	975	5	0.5128205128205128
46	2020-02-24	529	158	29.867674858223065
47	2020-02-25	841	79	9.3935790725327
48	2020-02-26	861	62	7.200929152148665
49	2020-02-27	1082	35	3.234750462107209
50	2020-02-28	1378	62	4.499274310595065
51	2020-02-29	1766	56	3.1710079275198186
52	2020-03-01	1799	46	2.5569760978321288
53	2020-03-02	1671	55	3.291442250149611
54	2020-03-03	2315	81	3.4989200863930883
55	2020-03-04	2981	91	3.0526668903052667
56	2020-03-05	2505	88	3.5129740518962076
57	2020-03-06	3648	106	2.905701754385965
58	2020-03-07	3681	114	3.0969845150774247
59	2020-03-08	3726	134	3.5963499731615673

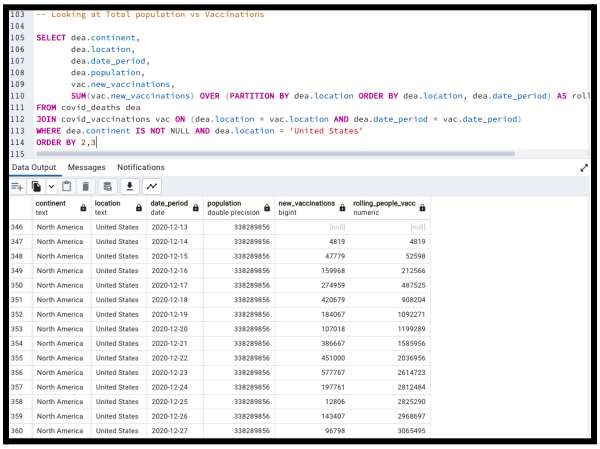
[Step 8's Results(Global List)]

```sql
-- Looking at overall total cases, deaths, and death percentage to date.
SELECT --date_period,
       SUM(new_cases) AS total_new_cases,
       SUM(new_deaths) AS total_new_deaths,
       SUM(new_deaths)/SUM(new_cases)*100 AS death_percentage
       --total_deaths,
       --(total_deaths/total_cases)*100 AS death_percentage
FROM covid_deaths
--WHERE location = 'United States'
WHERE continent IS NOT NULL
--GROUP BY 1
ORDER BY 1
```

Data Output | Messages | Notifications

	total_new_cases double precision	total_new_deaths double precision	death_percentage double precision
1	759432866	6869457	0.9045509231358443

[Step 9's Results]



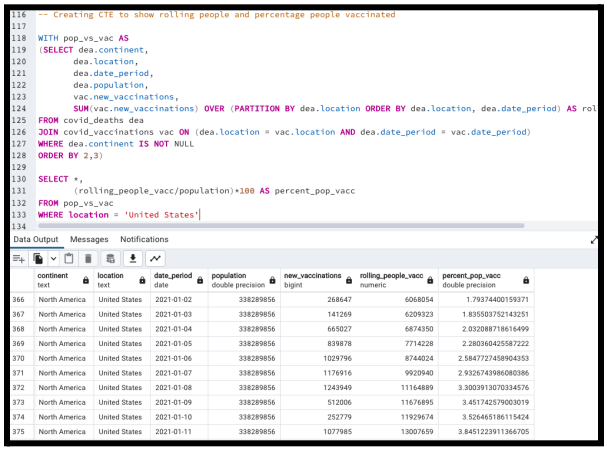
```
-- Looking at Total population vs Vaccinations

SELECT dea.continent,
       dea.location,
       dea.date_period,
       dea.population,
       vac.new_vaccinations,
       SUM(vac.new_vaccinations) OVER (PARTITION BY dea.location ORDER BY dea.location, dea.date_period) AS roll
FROM covid_deaths dea
JOIN covid_vaccinations vac ON (dea.location = vac.location AND dea.date_period = vac.date_period)
WHERE dea.continent IS NOT NULL AND dea.location = 'United States'
ORDER BY 2,3
```

Data Output | Messages | Notifications

	continent text	location text	date_period date	population double precision	new_vaccinations bigint	rolling_people_vacc numeric
346	North America	United States	2020-12-13	338289856	[null]	[null]
347	North America	United States	2020-12-14	338289856	4819	4819
348	North America	United States	2020-12-15	338289856	47779	52598
349	North America	United States	2020-12-16	338289856	159968	212566
350	North America	United States	2020-12-17	338289856	274959	487525
351	North America	United States	2020-12-18	338289856	420679	908204
352	North America	United States	2020-12-19	338289856	184067	1092271
353	North America	United States	2020-12-20	338289856	107018	1199289
354	North America	United States	2020-12-21	338289856	386667	1585956
355	North America	United States	2020-12-22	338289856	451000	2036956
356	North America	United States	2020-12-23	338289856	577767	2614723
357	North America	United States	2020-12-24	338289856	197761	2812484
358	North America	United States	2020-12-25	338289856	12806	2825290
359	North America	United States	2020-12-26	338289856	143407	2968697
360	North America	United States	2020-12-27	338289856	96798	3065495

[Step 10's Results]



```
-- Creating CTE to show rolling people and percentage people vaccinated

WITH pop_vs_vac AS
(SELECT dea.continent,
       dea.location,
       dea.date_period,
       dea.population,
       vac.new_vaccinations,
       SUM(vac.new_vaccinations) OVER (PARTITION BY dea.location ORDER BY dea.location, dea.date_period) AS rol
FROM covid_deaths dea
JOIN covid_vaccinations vac ON (dea.location = vac.location AND dea.date_period = vac.date_period)
WHERE dea.continent IS NOT NULL
ORDER BY 2,3)

SELECT *,
       (rolling_people_vacc/population)*100 AS percent_pop_vacc
FROM pop_vs_vac
WHERE location = 'United States'
```

Data Output | Messages | Notifications

	continent text	location text	date_period date	population double precision	new_vaccinations bigint	rolling_people_vacc numeric	percent_pop_vacc double precision
366	North America	United States	2021-01-02	338289856	268647	6068054	1.793744001593 71
367	North America	United States	2021-01-03	338289856	141269	6209323	1.835503752143251
368	North America	United States	2021-01-04	338289856	665027	6874350	2.032088718616499
369	North America	United States	2021-01-05	338289856	839878	7714228	2.280360425587222
370	North America	United States	2021-01-06	338289856	1029796	8744024	2.584772745890435 3
371	North America	United States	2021-01-07	338289856	1176916	9920940	2.932674398608038 6
372	North America	United States	2021-01-08	338289856	1243949	11164889	3.300391307033457 6
373	North America	United States	2021-01-09	338289856	512006	11676895	3.451742579003019
374	North America	United States	2021-01-10	338289856	252779	11929674	3.526465186115424
375	North America	United States	2021-01-11	338289856	1077985	13007659	3.845122391136 6705

[Step 11's Results]

```
-- Creating CTE to show rolling people and percentage people dead

WITH pop_vs_death AS
(SELECT dea.continent,
       dea.location,
       dea.date_period,
       dea.population,
       dea.new_deaths,
       SUM(dea.new_deaths) OVER (PARTITION BY dea.location ORDER BY dea.location, dea.date_period) AS rolling_p
FROM covid_deaths dea
JOIN covid_vaccinations vac ON (dea.location = vac.location AND dea.date_period = vac.date_period)
WHERE dea.continent IS NOT NULL
ORDER BY 2,3)

SELECT *,
       (rolling_people_dead/population)*100 AS percent_pop_dead
FROM pop_vs_death
WHERE location = 'United States'
```

Data Output | Messages | Notifications

	continent text	location text	date_period date	population double precision	new_deaths double precision	rolling_people_dead double precision	percent_pop_dead double precision
169	North America	United States	2020-06-19	338289856	684	119739	0.035395385902437466
170	North America	United States	2020-06-20	338289856	643	120382	0.0355854595888326
171	North America	United States	2020-06-21	338289856	755	121137	0.035808640977990185
172	North America	United States	2020-06-22	338289856	576	121713	0.03597890916362565
173	North America	United States	2020-06-23	338289856	413	122126	0.03610099381756218
174	North America	United States	2020-06-24	338289856	415	122541	0.03622366968047662
175	North America	United States	2020-06-25	338289856	746	123287	0.03644419062923365
176	North America	United States	2020-06-26	338289856	692	123979	0.036648748935587 36
177	North America	United States	2020-06-27	338289856	2533	126512	0.03739751510609884
178	North America	United States	2020-06-28	338289856	706	127218	0.03760621187529785

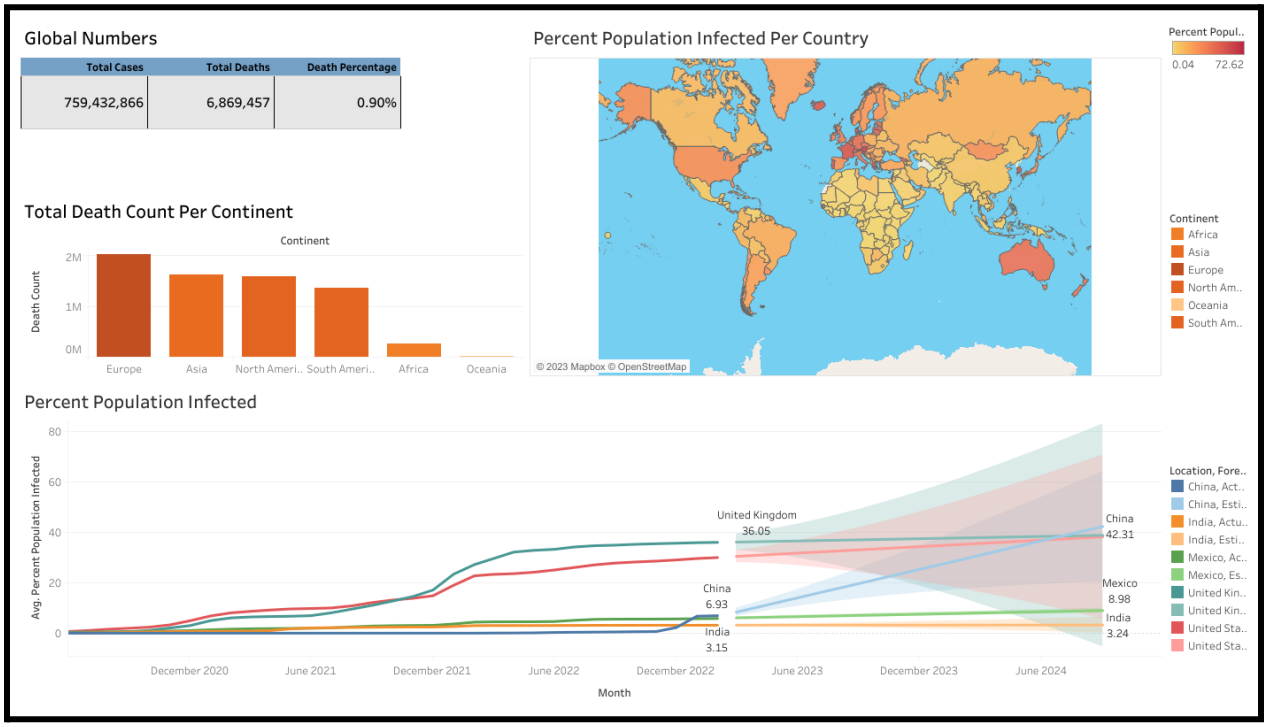
Interpreting statistical findings:

1. Our project requires us to mainly use the following columns: **continent, location, population, date_period, total_cases, new_cases, total_deaths, new_deaths, and new_vaccinations.**
2. By living in the USA, you would have had a **1.83%** chance of death in Feb. 2021, **1.22%** chance of death in Feb. 2022, and a **1.08%** chance of death in Feb. 2023.
3. By 2023, **30.22% of Americans had been infected** by COVID-19. The low percentage of infected Americans could be due to the USA's low population density.
4. The top 10 countries for **infection rate ranged from 72.62% - 59.92%**. All of these countries had populations less than 8.9 million, so there may be a relationship between small populations and infection rate.
5. The top 10 countries for **death count ranged from 1,111,342 - 161,397**. The highly populated countries were representative of this list with the USA leading the way.
6. The continents with the highest **death count in descending order** are: Europe, Asia, North America, South America, Africa, and Oceania.
7. By living on Earth, you would have had a **3.17%** chance of death in Feb. 29th, 2020, **2.15%** chance of death in Feb. 28th, 2021, **0.54%** chance of death in Feb. 28th 2022, and a **0.55%** chance of death in Feb. 28th, 2023.
8. Summarizing the global total cases, total deaths and death percentage as: **759,432,866 total cases** of COVID-19, **6,869,457 deaths**, and a **0.90% death percentage**.
9. Summarizing the total vaccinations vs. the USA's population. On Dec. 14th, 2020 **4,819 were vaccinated**, on Dec. 14th, 2021 **498,987,665 were vaccinated**, and on Dec. 14th, 2022 **662,646,858 were vaccinated**.
10. Explaining the vaccination percentage of the USA on a per day basis. On Dec. 28th, 2020 **1.08% were vaccinated**, on Dec. 28th, 2021 **152.97% were vaccinated**, and on Dec. 28th, 2022 **196.76% were vaccinated**.
11. A daily account of the total deaths, rolling count and deaths as a percentage of the USA's population: On Mar. 22nd, 2020 a **rolling death count of 374** and **death percentage of 0.00011%**, on Mar. 22nd, 2021 a **rolling death count of 545,840** and **death percentage of 0.16%**, and on Mar. 22nd, 2022 a **rolling death count of 971,478** and **death percentage of 0.28%**.

Step 5: Share

In this step, we are creating visualizations and communicating our findings based on our analysis.

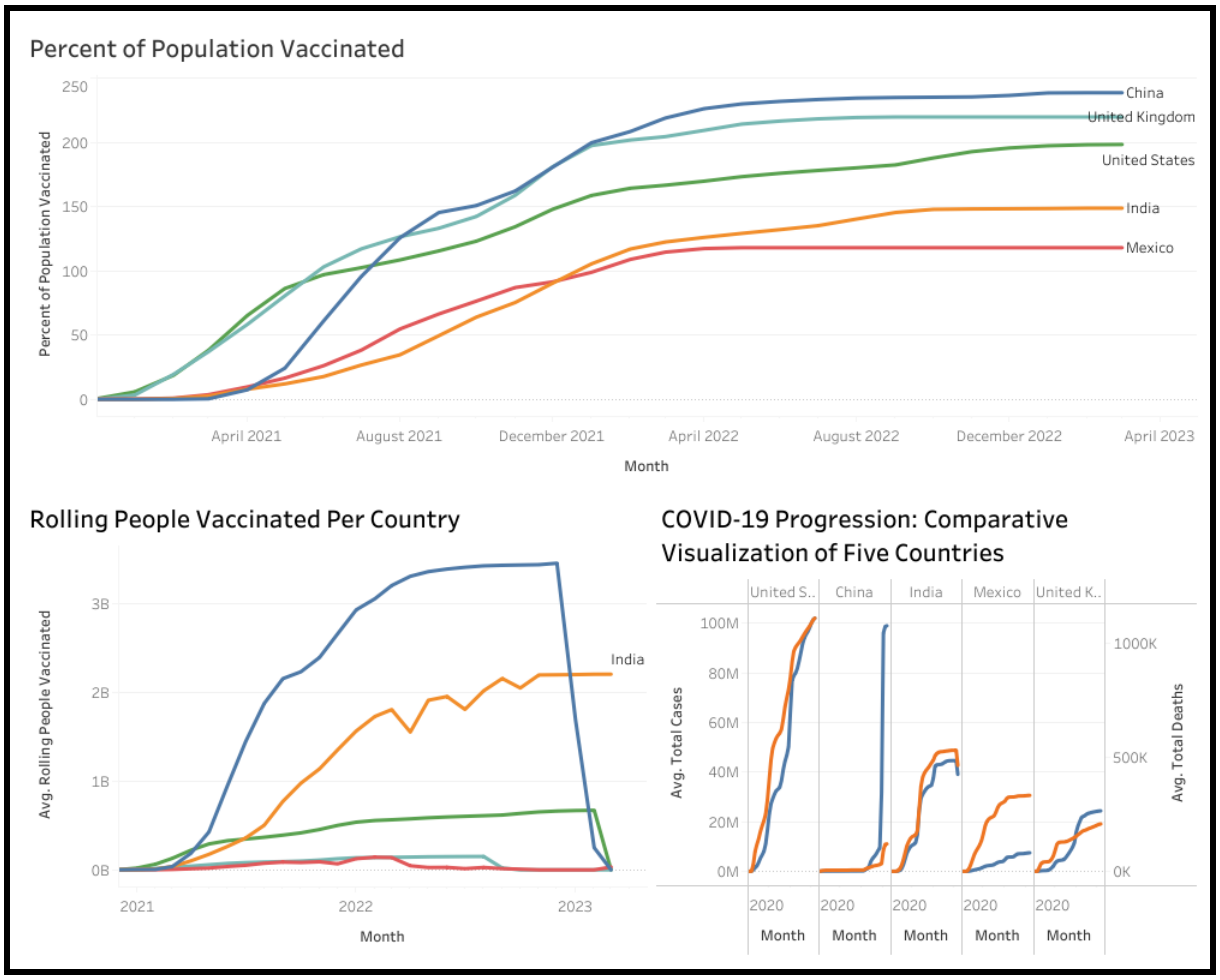
5.1 Data Visualizations and Findings



Looking at Global Deaths & Infections

This dashboard offers a global perspective on COVID-19 data, providing in-depth data points for total cases, deaths, and death percentage, with a country-level breakdown. We aim to highlight the total number of deaths across continents as well as the percentage of infected population in select countries.

1. As of March 2023, we discovered the average person has a 0.90% chance of dying from COVID-19.
2. With global deaths at 6,869,457, we uncovered European deaths leading the way, accounting for 29.45% of the total.
3. Among our select countries, the United Kingdom and United States led with 36.05% and 30.01% of their populations infected, with China forecasted to overtake them at 42.31% in late 2024. China relaxing Covid policies may be the cause.



Vaccination & Death Statistics For Select Countries

In this dashboard, we are looking at the rolling count for people vaccinated, percent of the population vaccinated, and comparing average total deaths to average total cases.

1. We discovered China led the vaccination drive with 239.0% of its population vaccinated, followed by the U.K. at 220.1% and the U.S. at 198.7%. China's centralized political system could be a reason for the higher percentage.
2. Since late 2022, we noticed vaccination drives have started to decelerate significantly measured by the average rolling count of people vaccinated. This may be because the world has relaxed covid policies and people are familiarized with Covid-19.
3. Among the selected countries, the USA, India, and Mexico lead the way for total cases to total deaths. This may be caused by their political systems which favor individual freedoms.

. . .

Step 6: Act

In the final step, we will be delivering our insights and providing recommendations based on our analysis.

Here, we revisit our business questions and share with you our high-level business recommendations.

1. **What is the global impact?**
    - The total cases and deaths from COVID-19 are staggering, and it is clear that the pandemic has had a major impact on people around the world. Consequently, it is important to continue supporting efforts to counter the pandemic globally, including through vaccine distribution and other measures.

2. **How do different countries fare when it comes to infection rate?**
    - The percentage of Americans who have been infected is lower than in some other countries, but this may be caused by the country's low population density. Therefore, it is crucial to know how population density and other local factors may be affecting the spread of the virus in communities relevant to you.

3. **What are the trends identified?**
    - While the risk of death has decreased in the USA over the past year, the rolling death count and death percentage continues to rise. As a result, it is necessary to take a long-term view by continuing to assess your local situation to understand if efforts to address the pandemic are effective over time.