GLOBAL COVID DEATHS:
SQL ANALYSIS & TABLEAU VISUALIZATION

ASK STAGE

This is a data analysis portfolio project to demonstrate how raw data is queried and formatted to understand the global impact of the coronavirus pandemic during its peak. The dataset is both large in scope and globally diverse (and can be found here). The subsets in question reflect data collected from January 1, 2020, to April 30, 2021.

A word on data integrity. In terms of bias and credibility, both data sources used comply with ROCCC standards:

- **Reliable and original:** this is public data that contains accurate and unbiased (and ongoing) info on global COVID-19 statistics from 2020 to mid-2021.
- **Comprehensive and current:** this source contains exhaustive data needed to understand the impact of the pandemic during an 18-month peak across the world. While it is not current as of this project, it is a sample time window that is useful for the exploratory analysis at hand.
- **Cited:** infection and vaccination sources are publicly available data provided by Our World in Data, which obtains it's information from the World Health Organization, and is completely open access under the Creative Commons BY license.

For this project, a subset of the available data for both infections and vaccinations were studied and compared to derive patterns and draw compelling conclusions.

This project was inspired by Alex the Analyst. However, I did not use SQL Server Manager as he did in his presentation; I utilized BigQuery for this analysis, therefore some of the queries will have superficial differences, albeit the same execution.

PREPARE STAGE

In preparing the data, I converted the XLSX files to CSV format so that they would be compatible with BigQuery. These are large datasets, so uploading locally was not ideal. Instead, I uploaded them to Google Cloud Storage, from where they could be transferred into the BigQuery platform.

ANALYSIS STAGE

This dataset is ideal in scope and breadth to perform various SQL queries to derive insights. This is all about exploration, so we aren't looking at specific questions, but rather collecting intel about general patterns and trends. If any are discovered, we can compare statistics across different countries.

Skills used throughout this project: Joins, CTE's, Temp Tables, Windows Functions, Aggregate Functions, Creating Views, Converting Data Types.

```
Select *
from `sql-data-project-379900.Portfolio_Project.CovidDeaths`
Where continent is not null
order by 3,4;

Select *
from `sql-data-project-379900.Portfolio_Project.CovidVaccinations`
Where continent is not null
order by 3,4;
```

SELECTING DATA TO START WITH.

```
Select Location, date, total_cases, new_cases, total_deaths, population
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
Where continent is not null
order by 1,2;
```

EXAMINING TOTAL CASES VS TOTAL DEATHS. THIS SHOWS THE LIKELIHOOD OF DYING IF YOU CONTRACT COVID IN YOUR NATION.

```
Select Location, date, total_cases,total_deaths, (total_deaths/total_cases)*100 as
DeathPercentage
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
order by 1,2;
```

EXAMINING TOTAL CASES VS POPULATION. THIS SHOWS WHAT PERCENTAGE OF THE POP IS COVID POSITIVE.

```
Select Location, date, total_cases,total_deaths, (total_deaths/total_cases)*100 as
DeathPercentage
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
Where location LIKE '%United States%'
order by 1,2;
```

EXAMINING COUNTRIES WITH THE HIGHEST INFECTION RATE, COMPARED TO POPULATION.

```
Select Location, date, Population, total_cases,  (total_cases/population)*100 as
PercentPopulationInfected
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
Where location LIKE '%United States%'
order by 1,2;
```

WHICH COUNTRIES HAVE THE HIGHEST DEATH COUNT PER POPULATION.

```sql
Select Location, Population, MAX(total_cases) as
HighestInfectionCount,  Max((total_cases/population))*100 as PercentPopulationInfected
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
--Where location LIKE '%United States%'
Group by Location, Population
order by PercentPopulationInfected desc;
Select Location, MAX(cast(Total_deaths as int)) as TotalDeathCount
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
--Where location LIKE '%United States%'
Where continent is not null
Group by Location
order by TotalDeathCount desc;
```

BREAKING DATA DOWN BY CONTINENT.
SHOWING CONTINENTS WITH THE HIGHEST DEATH COUNT PER POPULATION.

```sql
Select continent, MAX(cast(Total_deaths as int)) as TotalDeathCount
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
--Where location LIKE '%United States%'
Where continent is not null
Group by continent
order by TotalDeathCount desc;
```

(More accurate code, not for visualization purposes)

```sql
Select location, MAX(cast(Total_deaths as int)) as TotalDeathCount
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
--Where location LIKE '%United States%'
Where continent is null
Group by location
order by TotalDeathCount desc;
```

QUERYING GLOBAL NUMBERS.

```sql
Select SUM(new_cases) as total_cases, SUM(cast(new_deaths as int)) as total_deaths,
SUM(cast(new_deaths as int))/SUM(New_Cases)*100 as DeathPercentage
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
--Where location LIKE '%United States%'
where continent is not null
--Group By date
order by 1,2;
```

SELECTING DATASET TO VIEW TOTAL POPULATION WITH VACCINATIONS.

```sql
select *
From `sql-data-project-379900.Portfolio_Project.CovidVaccinations`;
```

PERCENTAGE OF POP THAT HAS AT LEAST ONE VACCINATION.

```sql
select *
From `sql-data-project-379900.Portfolio_Project.CovidDeaths` dea
Join `sql-data-project-379900.Portfolio_Project.CovidVaccinations` vac
  On dea.location = vac.location
  and dea.date = vac.date;

Select dea.continent, dea.location, dea.date, dea.population, vac.new_vaccinations
From `sql-data-project-379900.Portfolio_Project.CovidDeaths` dea
Join `sql-data-project-379900.Portfolio_Project.CovidVaccinations` vac
  On dea.location = vac.location
  and dea.date = vac.date
Where dea.continent is not null
Order by 1,2,3;
```

LOOKING AT TOTAL POP VS VACCINES
SHOWS PERCENTAGE OF POP THAT HAS RECEIVED AT LEAST ONE COVID VACCINE.

```sql
Select dea.continent, dea.location, dea.date, dea.population, vac.new_vaccinations
, SUM(Cast(vac.new_vaccinations as int)) OVER (Partition by dea.Location)
From `sql-data-project-379900.Portfolio_Project.CovidDeaths` dea
Join `sql-data-project-379900.Portfolio_Project.CovidVaccinations` vac
  On dea.location = vac.location
  and dea.date = vac.date
where dea.continent is not null
order by 2,3;


Select dea.continent, dea.location, dea.date, dea.population, vac.new_vaccinations
, SUM(Cast(vac.new_vaccinations as int)) OVER (Partition by dea.Location Order by
dea.location, dea.Date) as RollingPeopleVaccinated
--, (RollingPeopleVaccinated/population)*100
From `sql-data-project-379900.Portfolio_Project.CovidDeaths` dea
Join `sql-data-project-379900.Portfolio_Project.CovidVaccinations` vac
  On dea.location = vac.location
  and dea.date = vac.date
where dea.continent is not null
--order by 2,3;
```

USING CTS TO PERFORM CALCULATION ON PARTITION BY IN PREVIOUS QUERY.

```sql
with PopvsVac as (
  Select dea.continent, dea.location, dea.date, dea.population, vac.new_vaccinations
, SUM(Cast(vac.new_vaccinations as int)) OVER (Partition by dea.Location Order by
dea.location, dea.Date) as RollingPeopleVaccinated
--, (RollingPeopleVaccinated/population)*100
From `sql-data-project-379900.Portfolio_Project.CovidDeaths` dea
Join `sql-data-project-379900.Portfolio_Project.CovidVaccinations` vac
  On dea.location = vac.location
  and dea.date = vac.date
where dea.continent is not null
--order by 2,3;
```

```
)
select *
from PopvsVac;
```

USING A TEMP TABLE TO PERFORM CALCULATION ON PARTITION BY IN PREVIOUS QUERY.

```
CREATE TABLE `sql-data-project-379900.Portfolio_Project.PercentPopulationVaccinated`
(
continent string,
location string,
date datetime,
population int64,
new_vaccinations int64,
RollingPeopleVaccinated int64
);

INSERT INTO Portfolio_Project.PercentPopulationVaccinated
SELECT dea.continent, dea.location, dea.date, dea.population, vac.new_vaccinations
  , SUM(Cast(vac.new_vaccinations as int)) OVER (Partition by dea.Location Order by
dea.location, dea.Date) as RollingPeopleVaccinated
--, (RollingPeopleVaccinated/population)*100
From `sql-data-project-379900.Portfolio_Project.CovidDeaths` dea
Join `sql-data-project-379900.Portfolio_Project.CovidVaccinations` vac
  On dea.location = vac.location
  and dea.date = vac.date
where dea.continent is not null;
Select *, (RollingPeopleVaccinated/Population)*100 as vac_rate
From Portfolio_Project.PercentPopulationVaccinated;
```

CREATING A VIEW TO STORE FOR TABLEAU VISUALIZATIONS IN SECOND PART OF PROJECT.

```
CREATE VIEW Portfolio_Project.PercentPopulationVaccinated AS
SELECT dea.continent, dea.location, dea.date, dea.population, vac.new_vaccinations
  , SUM(Cast(vac.new_vaccinations as int)) OVER (Partition by dea.Location Order by
dea.location, dea.Date) as RollingPeopleVaccinated
--, (RollingPeopleVaccinated/population)*100
From `sql-data-project-379900.Portfolio_Project.CovidDeaths` dea
Join `sql-data-project-379900.Portfolio_Project.CovidVaccinations` vac
  On dea.location = vac.location
  and dea.date = vac.date
where dea.continent is not null;
```

RUN CODE FOR TABLEAU VISUALIZATIONS
NOTE: Instead of copying to a spreadsheet, in Bigquery just click save results and it will extract as a pre-made CVS file you can save as an Excel workbook.

```
Select SUM(new_cases) as total_cases, SUM(cast(new_deaths as int)) as total_deaths,
SUM(cast(new_deaths as int))/SUM(New_Cases)*100 as DeathPercentage
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
--Where location LIKE '%United States%'
```

```sql
where continent is not null
--Group By date
order by 1,2;

Select location, SUM(cast(new_deaths as int)) as TotalDeathCount
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
--Where location LIKE '%United States%'
Where continent is null
and location not in ('World', 'European Union', 'International')
Group by location
order by TotalDeathCount desc;

Select Location, Population, MAX(total_cases) as
HighestInfectionCount,  Max((total_cases/population))*100 as PercentPopulationInfected
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
--Where location LIKE '%United States%'
Group by Location, Population
order by PercentPopulationInfected desc;

Select Location, Population, date, MAX(total_cases) as
HighestInfectionCount,  Max((total_cases/population))*100 as PercentPopulationInfected
From `sql-data-project-379900.Portfolio_Project.CovidDeaths`
--Where location LIKE '%United States%'
Group by Location, Population, date
order by PercentPopulationInfected desc;
```

NOTE: In cleaning the fourth Excel doc, replace blanks with zeroes by filtering and copy/paste. Far more efficient than find/replace.

CONCLUSIONS, OBSERVATIONS, & ASSUMPTIONS

According to the current analysis, we can make some interesting observations, particularly about the United States.

- First observation: by the end of the timeline, America had about 10% infections compared to Canada's 3%.
- Second observation: out of every country on earth, America had the highest death count per population.
- Third observation: out of each continent on earth, North America had the highest death count per population.

Conclusion: America saw a significant infection and death toll compared to other countries during the height of the coronavirus pandemic.

SQL is not a visual medium for sharing results with an audience, so let's turn our attention here, in which I present key findings in Tableau Public via a visual dashboard. Included are global number counts; a bar chart depicting total deaths per continent; a map showing the percentage of

the population infected by country; and a timeline trend chart showing comparisons of infected populations between sample countries.