COVID-19 Data Analysis Project

What This Project Does

This project digs into COVID-19 data to understand the pandemic's impact across different countries and continents. I built this to answer questions like "What's the death rate?", "Which countries were hit hardest?", and "How did vaccination rates compare globally?"

The Data

The analysis uses two main tables:

- **CovidDeaths -** Contains daily case counts, deaths, and population data
- **CovidVaccinations -** Tracks testing and vaccination rollout across countries

Key Analyses

Understanding the Basics

I started by looking at the fundamentals - tracking total cases and deaths over time for each location. For India specifically, I calculated the death percentage to understand the risk if someone contracted COVID-19.

Infection Rates

One of the more interesting parts was comparing infection rates relative to population size. This helps identify which countries had the highest percentage of their population infected, which is more telling than just raw numbers.

Death Counts

I analyzed death counts both by country and by continent. The continent-level analysis required some data cleaning since the original dataset had some quirks in how it categorized locations.

Vaccination Coverage

The vaccination analysis looks at:

- Which countries have complete vaccination data
- Total number of people fully vaccinated
- Countries with missing vaccination data (surprisingly common)

Rolling Vaccination Rates

One of the more complex queries tracks the rolling count of vaccinated people over time using window functions. This shows how vaccination campaigns progressed day-by-day in each location.

Techniques Used

Data Aggregation: Lots of GROUP BY operations to summarize data at country and continent levels

Window Functions: Used PARTITION BY with running totals to track cumulative vaccinations

Joins: Combined the deaths and vaccinations tables to see the full picture

CTEs and Temp Tables: Created reusable query components for calculating vaccination percentages - showed this in a couple different ways to demonstrate different approaches

Views: Set up a view for the vaccination percentage data so it's easy to pull into visualization tools later

What I Learned

This project really drove home how important it is to:

- Check your data types (had to CAST some fields to integers)
- Handle NULL values carefully (especially with continent vs. location filtering)
- Think about what you're actually measuring (deaths per population vs. death rate among infected are very different things)

Running the Queries

These queries were written for SQL Server (you can see the `PortfolioProject..` database references). If you're using a different SQL flavor, you might need to adjust:

- The CONVERT/CAST syntax
- How window functions are written
- Temp table creation syntax

Future Ideas

Some things I'd like to add:

- Time-series analysis to identify waves/peaks
- Correlation between vaccination rates and death rates
- More granular age-group analysis if that data becomes available
- Comparison of different vaccine types and their rollout timing

A Quick Note

The data landscape around COVID-19 was messy - different countries reported things differently, testing availability varied wildly, and there were definitely data quality issues. These queries work with what's available, but real-world analysis would need more data validation and cleaning.