December 6, 2020

ETL Project Report

# Initial Post

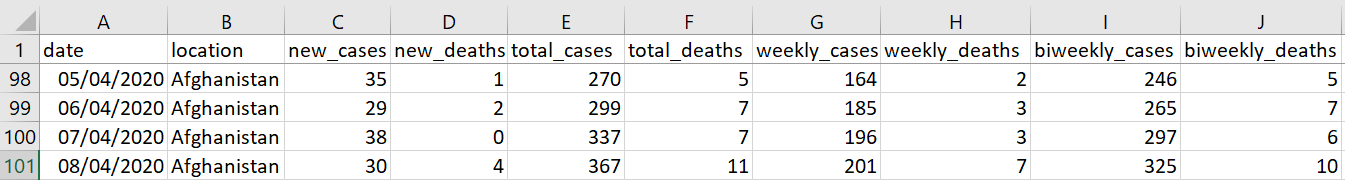
The final project proposal attempted to answer the following question: “How have countries around the world been impacted by three recent pandemics - SARS, H1N1, and COVID-19?” We decided to find the total number of cases and the total number of deaths for each country in the world for COVID-19, H1N1 (Influenza), and SARS. We found datasets for these diseases from different sources (linked in the section below) and combined them in SSIS.

# Description of Data Sources

We used three data sources for this project.

**Our World in Data COVID Full Data**

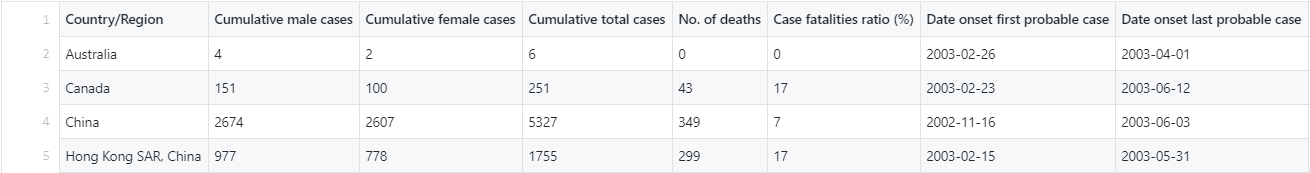
Source: <https://covid.ourworldindata.org/data/ecdc/full_data.csv>

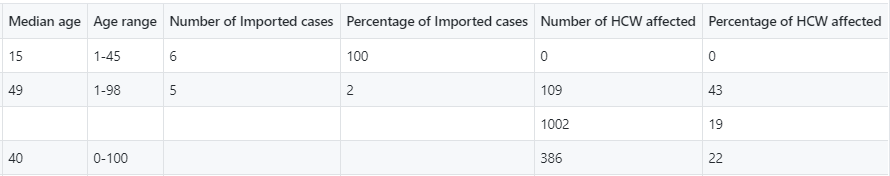


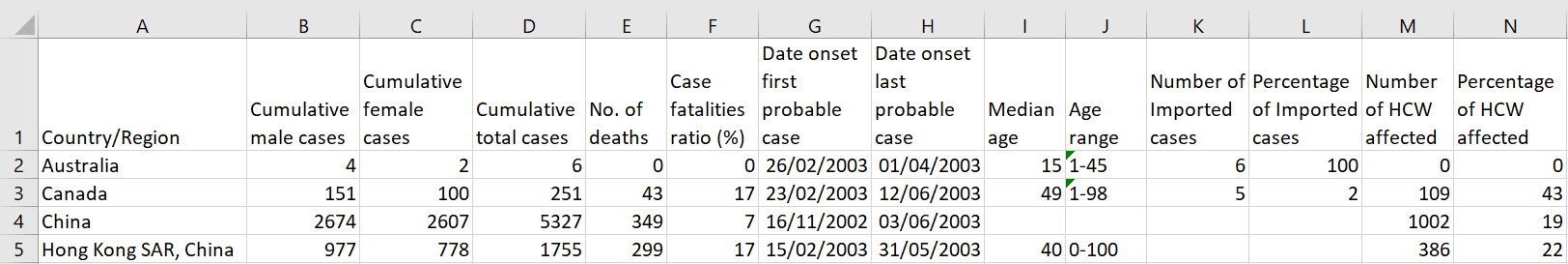
The Our World in COVID Full Data dataset includes data from 213 countries as well as a world total and international category (which most likely references COVID-19 cases on cruise ships and other international locations). The data for most countries begins on December 31st, 2019, although some countries’ data does not begin until a few months later, and continues until November 27th, 2020 in most cases, although some countries only have data until November 25th, 2020. The data for each country shows the number of new cases and new deaths for each day as well as the total number of cases and deaths in that country as of that date. Further, the data includes the weekly and biweekly number of deaths and cases for each date. We took this data set and kept only data from the most recent date on record for each country. We also only kept the data, location (country), total cases, and total deaths columns.

**SARS 2003 Outbreak Data**

Source: <https://github.com/imdevskp/sars-2003-outbreak-data-webscraping-code/blob/master/summary_data_clean.csv>







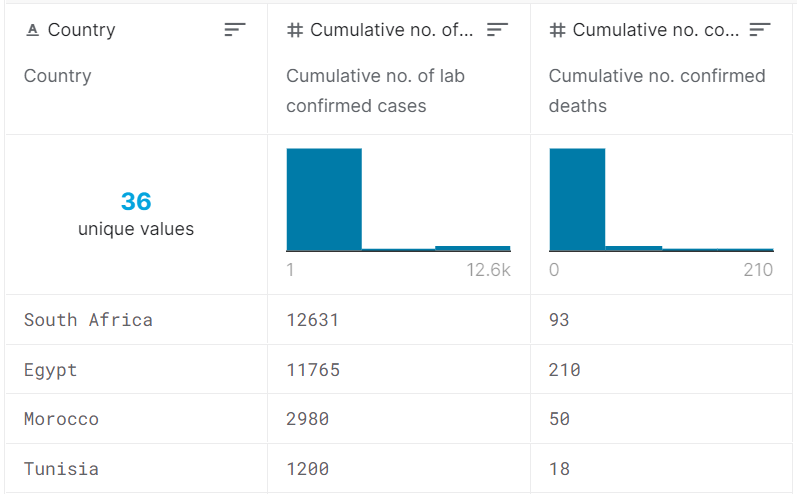
Severe Acute Respiratory Syndrome (SARS) is a viral disease that affects the respiration or the organs of respiration. This disease is caused by a SARS-associated coronavirus. It was first identified at the end of February 2003 throughout an outbreak that started in China and spread to four other countries.

Many of the patients identified with SARS were formerly healthy adults aged 25–70 years. A small number of suspected cases of SARS have been reported among children under 15 years. The case mortality among people with illness meeting the current WHO case definition for probable and suspected cases of SARS is around 3%.

This data is from 2003. The data for each country shows the number of cases between male and females and as well as the total number of cases and deaths in that country.

**H1N1 | 2009 | Swine Flu Pandemic**

Source: <https://www.kaggle.com/imdevskp/h1n1-swine-flu-2009-pandemic-dataset>



Multiple summary datasets were retrieved from this Kaggle project, which provide summaries of the number of cases and deaths for countries as of November 2009. The source for the data is Wikipedia. The following datasets were combined for our project’s use: summary\_africa.csv, summary\_asia.csv, summary\_europe.csv, summary\_south\_america.csv, and summary\_usa.csv. All the datasets were combined and deduplicated (some countries were present in both the Europe and Asia datasets). In addition, the dataset for the United States included state and territory level data so the total for all the United States was calculated and added to the combined H1N1 dataset.

The H1N1 dataset consists of the following columns: Country, Cumulative no. of lab confirmed cases, Cumulative no. confirmed deaths. Additional columns are present in some of the datasets but were out of scope for this project, so they were excluded.

# Justification of the Unified View on Data

For this project, we considered that it would be interesting to compare data from multiple pandemics, to especially reflect on the scope and impact of the COVID-19 pandemic.

A united view on the data from our three sources on the COVID-19 pandemic, the H1N1 (swine flu) pandemic, and the SARS pandemic is necessary in order to compare the magnitude of each pandemic by looking at the case fatality rate for a country affected by all three pandemics. A united view of this data is also necessary to compare the scope and reach of each pandemic by calculating how many countries have been affected by each disease. Finally, a united view of the data is needed to compare which countries were hardest hit by each pandemic by finding which countries had the highest total number of cases for each pandemic as well as which had the highest total number of deaths.

# Project Instructions

The project requires the following:

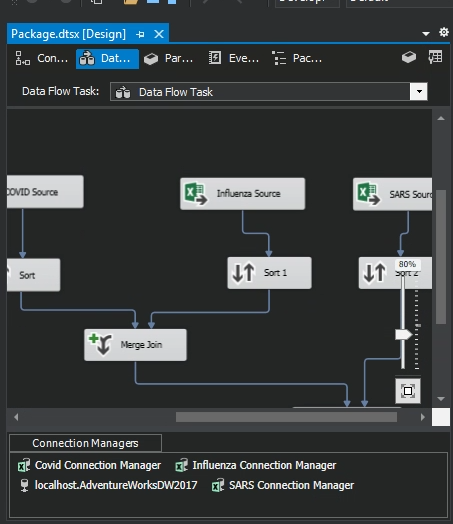
* Visual Studio with SQL Integration Services installed
* SQL Server Management Studio, with an AdventureWorks2017DW database connection
* 3 excel tables used as sources for this project

To begin this project, for the purposes of simplicity the tables can be put in a “tables” folder on the user’s Desktop. Visual Studio must be opened, and a new Integration Services Project must be selected. The first component to be brought over from the SSIS Toolbox is the Data Flow component.

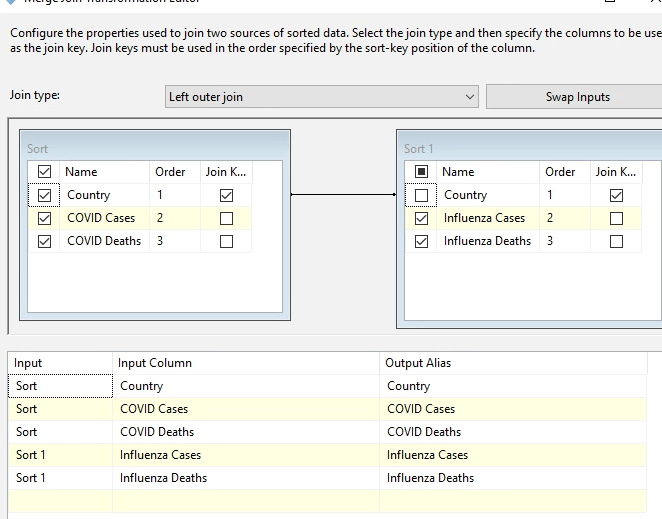
Double clicking on the data flow component brings us the Data Flow tab. From there, three excel sources can be brought from Sources in SSIS Toolbox. When double-clicked, each source has a Connection Manager option. New Connection Manager must be selected, and the path should direct to the corresponding table for each one. The format, if not automatically selected, should be Excel 2007-2010.

After finding the tables, we select Columns and edit the column names to be uniform. Any table should begin with “Countries”. Then, they should have “Covid Cases” or “Covid Deaths” – or whatever the corresponding outbreak is in the other tables. It is good to rename the tables and connections managers their corresponding names.

After the tables are added and Connection Managers configured, a Sort is brought over under each source. This must be done in order to allow for a merge join after. They should already have the expected column order.

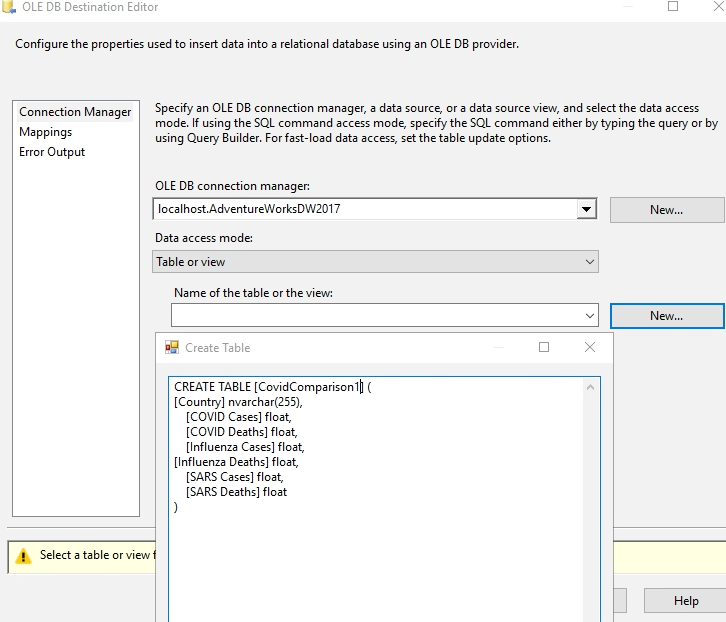


The following step is an important one, where the first join between the two tables happens, and then that join is merge with the third table. Merge Join is brought over from the SSIS toolbox and positioned between the first two sorts. It is important to select Join Key for “Countries”, and leave the checkmark on the left of the first “Countries” column selected, while only selecting “Join Key” for the other one:

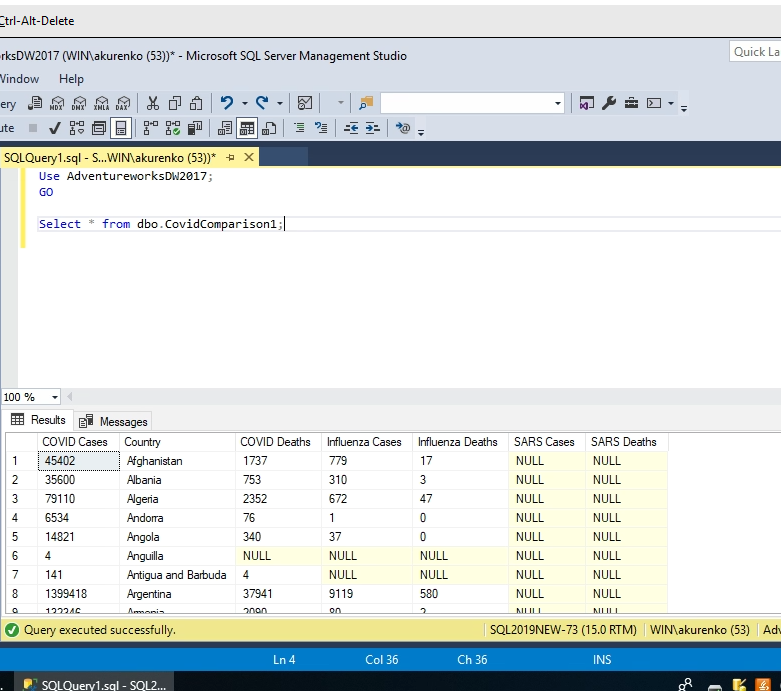


A similar process is undergone with the last table and the previous merge. Finally, and OLE DB Destination is brought from the toolkit. For our project we configured it to connect to the KSU VM SQL server to the localhost.AdventureworksDW2017 database.

We must select “New Table” and decide on a name for the table. After the Merge column positions may have moved, so it must be created in the order we want them.



We can also ensure the Mappings are correct, although they are usually by default. The project is ready to run. Clicking on Start Debugging brings us to our resulting table in SQL Server. Bringing up SQL Server. AdventureworksDW1027 database, we should see the table there. Every iteration of this project running makes a table, so when we remake the table, we must drop the existing one in SQL Server.



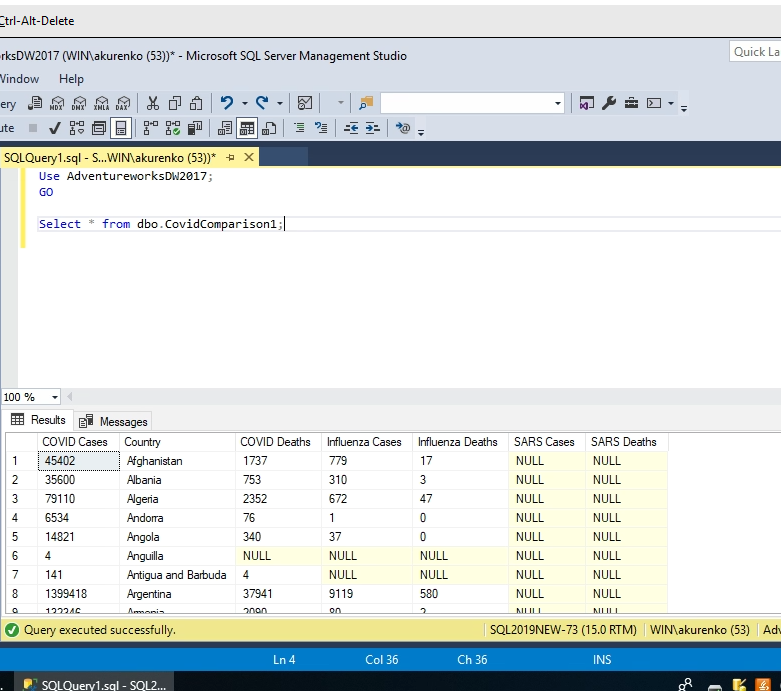
**Queries**

After loading the project which outputs the table to SQL, we open the AdventureworksDW2017, and perform the following query to retrieve all of the data.

Use AdventureworksDW2017;

GO

Select \* from dbo.CovidComparison1;



# Analysis

For this project, we sought to explore the following four questions.

**How many countries has each pandemic affected?**

This can be visualized by using a SQL query to count the number of countries where cases are not null for the disease. Comparing the number of countries affected shows the scope and reach of the disease.

SELECT COUNT(DISTINCT Country) as COVIDCountryCount

FROM dbo.CovidComparison1

WHERE [COVID cases] IS NOT NULL;

SELECT COUNT(DISTINCT Country) as InfluenzaCountryCount

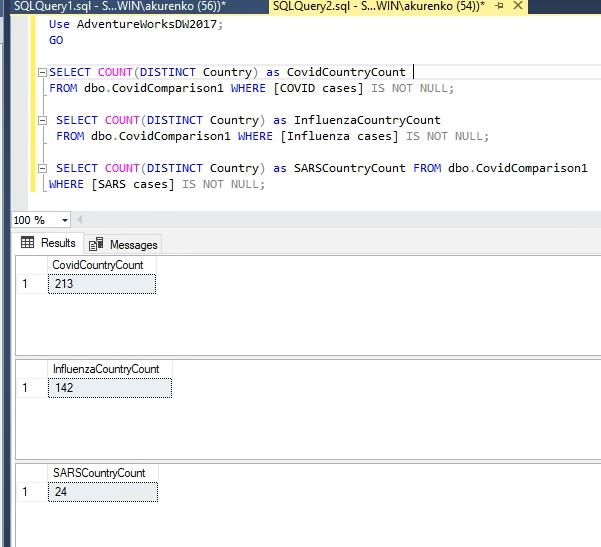
FROM dbo.CovidComparison1

WHERE [Influenza cases] IS NOT NULL;

SELECT COUNT(DISTINCT Country) as SARSCountryCount

FROM dbo.CovidComparison1

WHERE [SARS cases] IS NOT NULL;



**How many countries have been affected by all three pandemics?**

In addition, we would like to see which countries have data for each outbreak. To do this we query:

Use AdventureWorksDW2017;

GO

SELECT \*

FROM dbo.CovidComparison1

WHERE [Covid cases] != 0

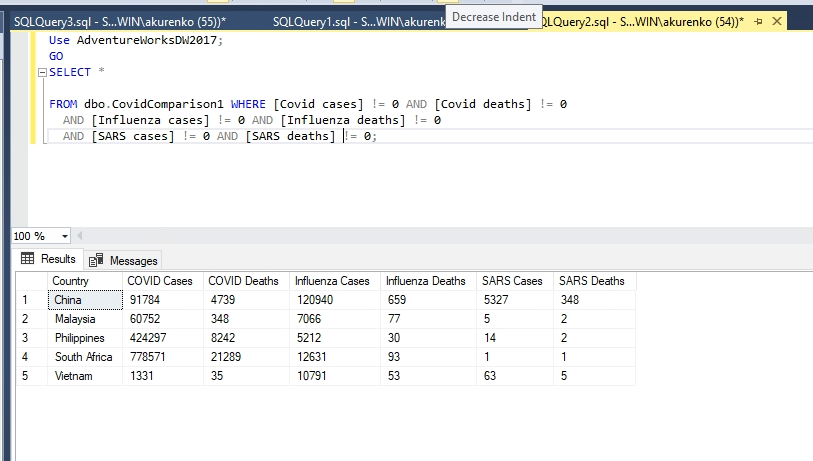
AND [Covid deaths] != 0

AND [Influenza cases] != 0

AND [Influenza deaths] != 0

AND [SARS cases] != 0

AND [SARS deaths] != 0;



The above query filtered out all value of 0 or NULL. We have also gathered some data for the highest number of COVID Deaths, highest number of Influenza Cases, and the highest number of Influenza Deaths:

Use AdventureworksDW2017;

GO

Select \* from dbo.CovidComparison1 WHERE [Covid Deaths] IS NOT NULL

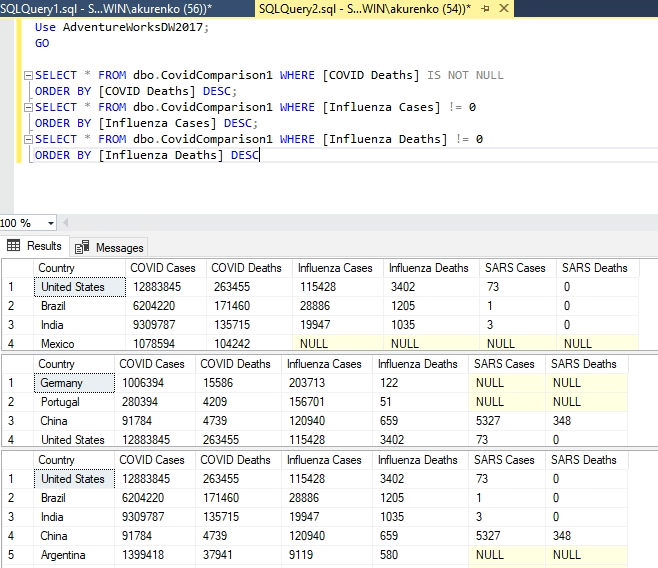
Order by [Covid Deaths] DESC;

Select \* from dbo.CovidComparison1 WHERE [Influenza Cases] IS NOT NULL

Order by [Influenza Cases] DESC;

Select \* from dbo.CovidComparison1 WHERE [Influenza Deaths] IS NOT NULL

Order by [Influenza Deaths] DESC;



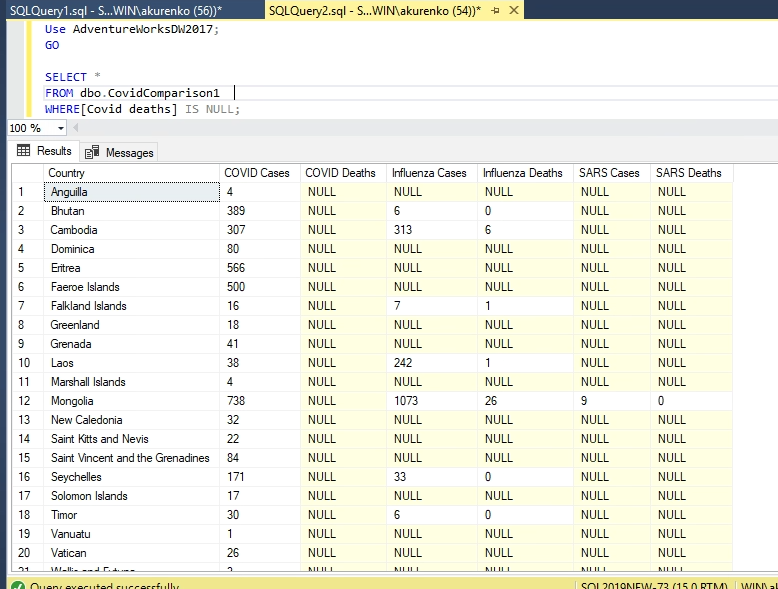
**How many countries have experienced no deaths during any pandemic?**

In addition, we can see countries with no deaths for COVID or any other outbreak:

SELECT \*

FROM dbo.CovidComparison1

WHERE [Covid deaths] IS NULL;



Or countries with no deaths from any outbreak:

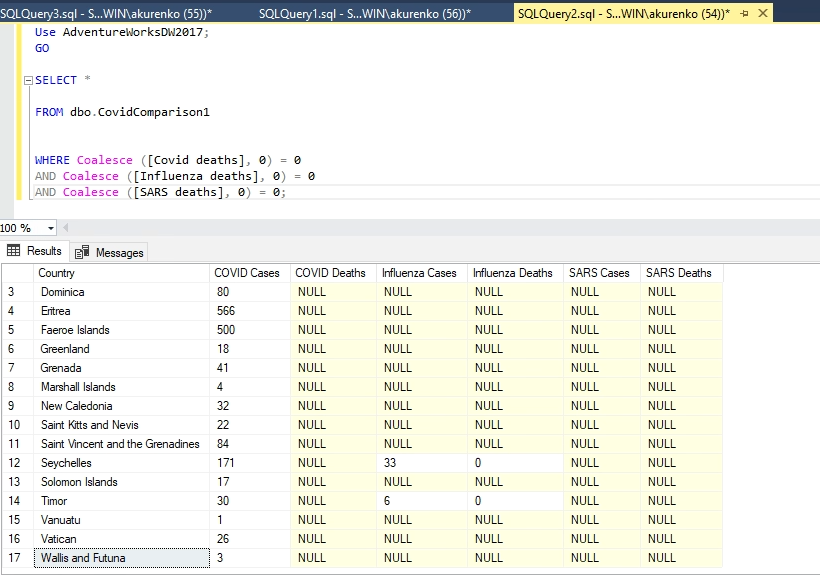
SELECT \*

FROM dbo.CovidComparison1

WHERE Coalesce ([Covid deaths], 0) = 0

AND Coalesce ([Influenza deaths], 0) = 0

AND Coalesce ([SARS deaths], 0) = 0;



The above query returns both null and 0 values.

**For a country that experienced more than two of the pandemics, how does the case fatality ratio (CFR) compare?**

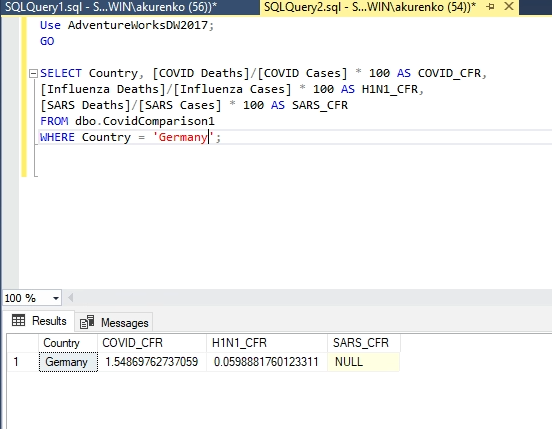
The case-fatality ratio or case fatality rate is a measure of disease severity and is the "proportion of individuals diagnosed with a disease who die from that disease”[[1]](#footnote-2)￼. CFR can be calculated by dividing the number of deaths by the number of cases and multiplying by 100 to get a percentage. The case fatality ratio is best calculated at the end of an outbreak and can be used to assess the deadliness of an outbreak. There is also a formula for ongoing epidemic or pandemic (i.e. COVID-19), but for the purposes of this report we will use the same formula for all three diseases.

This query calculates the CFR for all three pandemics for the country Germany.

SELECT Country, [COVID Deaths]/[COVID Cases] \* 100 AS COVID\_CFR, [Influenza Deaths]/[Influenza Cases] \* 100 AS H1N1\_CFR, [SARS Deaths]/[SARS Cases] \* 100 AS SARS\_CFR

FROM dbo.CovidComparison1

WHERE Country = 'Germany';

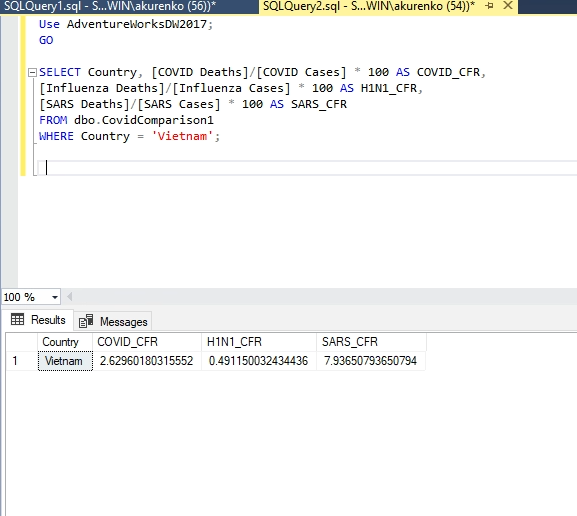


This query calculates the CFR for all three pandemics for the country Vietnam.

SELECT Country, [COVID Deaths]/[COVID Cases] \* 100 AS COVID\_CFR, [Influenza Deaths]/[Influenza Cases] \* 100 AS H1N1\_CFR, [SARS Deaths]/[SARS Cases] \* 100 AS SARS\_CFR

FROM dbo.CovidComparison1

WHERE Country = 'Vietnam';



**For each pandemic, which country had the highest total number of cases? Which had the highest total number of deaths? How do these numbers compare?**

This can be found by using an SQL query to find the row containing the highest value for the given column and repeating this for each of the total cases and total deaths columns. Finding which countries had the most cases and which had the most deaths for each pandemic shows which parts of the world were hit hardest by each pandemic. Comparing the numbers themselves from these countries demonstrates the impact and magnitude of these diseases in the countries that have been most affected by them. The SQL Query we used:

Use AdventureWorksDW2017;

GO

Declare @max\_COVID\_Cases int, @max\_Influenza\_Cases int, @max\_SARS\_Cases int,

@max\_COVID\_Deaths int, @max\_Influenza\_Deaths int, @max\_SARS\_Deaths int;

SELECT @max\_COVID\_Cases = MAX([COVID Cases]) FROM dbo.CovidComparison1;

SELECT @max\_COVID\_Deaths = MAX([COVID Deaths]) FROM dbo.CovidComparison1;

SELECT @max\_SARS\_Cases = MAX([SARS Cases]) FROM dbo.CovidComparison1;

SELECT @max\_SARS\_Deaths = MAX([SARS Deaths]) FROM dbo.CovidComparison1;

SELECT @max\_Influenza\_Cases = MAX([Influenza Cases]) FROM dbo.CovidComparison1;

SELECT @max\_Influenza\_Deaths= MAX([Influenza Deaths]) FROM dbo.CovidComparison1;

SELECT Country, [COVID Cases] FROM dbo.CovidComparison1

WHERE [COVID Cases] = @max\_COVID\_Cases;

SELECT Country, [SARS Cases] FROM dbo.CovidComparison1

WHERE [SARS Cases] = @max\_SARS\_Cases;

SELECT Country, [Influenza Cases] FROM dbo.CovidComparison1

WHERE [Influenza Cases] = @max\_Influenza\_Cases;

SELECT Country, [COVID Deaths] FROM dbo.CovidComparison1

WHERE [COVID Deaths] = @max\_COVID\_Deaths;

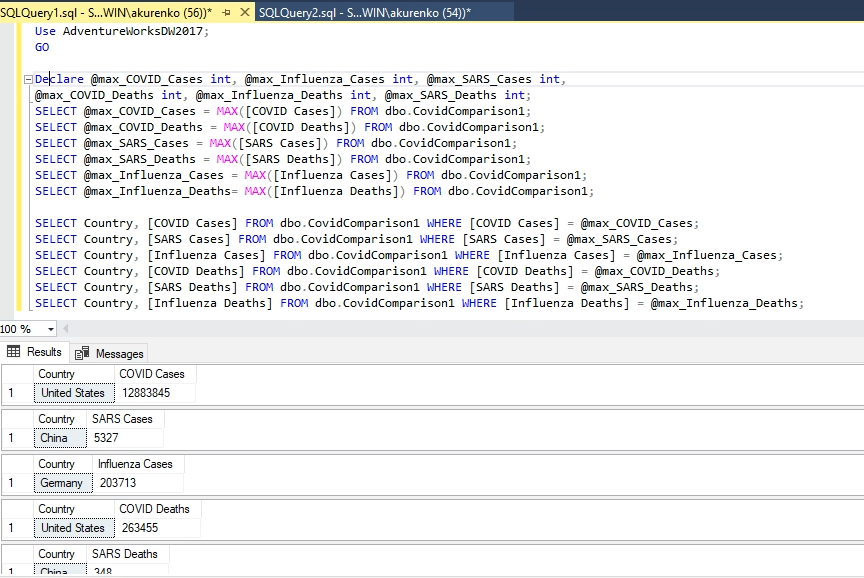
SELECT Country, [SARS Deaths] FROM dbo.CovidComparison1

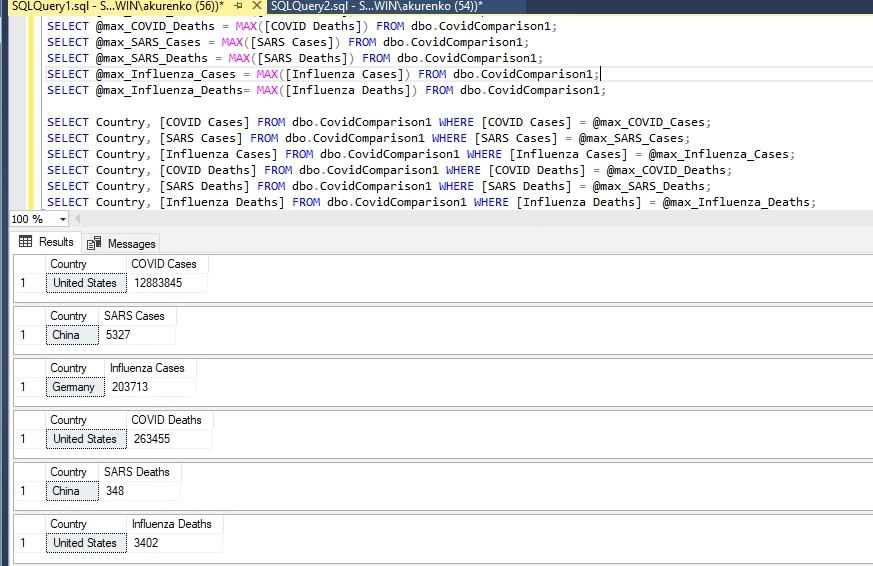
WHERE [SARS Deaths] = @max\_SARS\_Deaths;

SELECT Country, [Influenza Deaths] FROM dbo.CovidComparison1

WHERE [Influenza Deaths] = @max\_Influenza\_Deaths;

GO





Reflection

After exploring the project as a group, our report expanded significantly from our initial post on the discussion board. After some time, we were able to more fully explore the datasets and understand our focus on comparing the impact of multiple pandemics, including how many countries were affected by each pandemic, comparing the case fatality rate for each pandemic for a single country, and comparing the countries which had the highest total number of cases and deaths for each pandemic.

**How many countries has each pandemic affected?**

In the Analysis section, we expected to use a query to count the number of countries affected by each pandemic. As you can see based on the screenshots of the results in the Analysis section above, we successfully achieved the results we expected. We learned that of the three pandemics, SARS spread to the least number of countries (24 countries), and H1N1 (142 countries) and COVID-19 (213 countries) have spread to significantly more of the world.

**How many countries have been affected by all three pandemics?**

In the Analysis section, we expected to use a query to count the number of countries affected by all three pandemics. As you can see based on the screenshots of the results in the Analysis section above, we successfully achieved the results we expected. The results of the query showed that a relatively small number of countries were affected by all 3 pandemics. This could be due to the relatively low spread of Influenza and SARS compared to COVID-19, as well due to the significantly lower amount of country data provided for Influenza.

**How many countries have experienced no deaths during any pandemic?**

In addition, we could see which countries did not have any deaths for any outbreak or for Covid specifically. Inversely to the previous question, we found that low population as well as remoteness from other countries is a deciding factor here. Island nations such as Marshal Island and Greenland comprised most of those countries.

**For a country that experienced more than two of the pandemics, how does the case fatality ratio (CFR) compare?**

In the Analysis section, we expected to use a query to calculate the case fatality ratio for each pandemic for a country that experienced at least two of the pandemics. The results of the query show how deadly each pandemic was comparatively in Germany and Vietnam. While using CFR for COVID-19, as an ongoing pandemic, is not the most recommended method, comparing the results show H1N1 as the least deadly and SARS as the deadliest. COVID-19 is somewhere in the middle of those two.

**For each pandemic, which country had the highest total number of cases? Which had the highest total number of deaths? How do these numbers compare?**

In the Analysis section, we expected to use a query to find which country had the highest total number of cases and which had the highest total number of deaths for each pandemic. As you can see based on the screenshots of the results in the Analysis section above, we successfully achieved the results we expected.

The results of the query showed that the United States had the highest total number of COVID-19 cases with 12,883,845. Germany had the highest total number of H1N1 cases with 203,713. China had the highest total number of SARS cases with 5,327. The United States had the highest total number of COVID-19 deaths with 263,455. The United States had the highest total number of H1N1 deaths with 3,402. China had the highest total number of SARS deaths with 348.

We learned that, in the countries most affected by each pandemic, there were far many more cases of COVID-19 and deaths from COVID-19 than there were of the other two pandemics. We also saw that SARS had far fewer cases and deaths than the other two pandemics. We were surprised to learn that the United States had the highest number of deaths in two out of the three pandemics we looked at, COVID-19 and H1N1, and the highest number of COVID-19 cases. This could stem from the fact that the US has a large population, or it could signal that the US has handled these pandemics poorly, or a combination of the two. Unsurprisingly, China had both the highest number of SARS cases and the highest number of SARS deaths. This makes sense because the SARS pandemic was mostly geographically focused on China and the surrounding countries to a lesser extent. Finally, we learned that Germany had the highest number of H1N1 cases, but did not have the highest number of deaths, which could signal an effective healthcare system in the country, or a generally healthy and/or youthful population, or a combination of the two.

1. https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19 [↑](#footnote-ref-2)