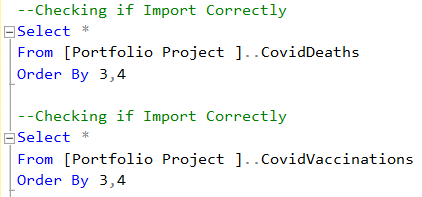
*Covid-19 Pandemic Case Study*

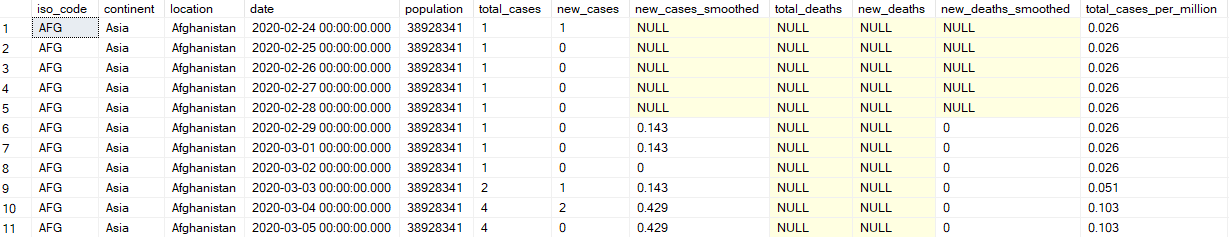
In this case study, I will demonstrate my proficiency utilizing SQL by performing exploratory data analysis. I will use a variety of functions in Microsoft SQL Server to clean, transform, wrangle, and manipulate the dataset. Afterwards, I will demonstrate my proficiency using Tableau to produce insightful and visually-appealing graphics in relation to the data.

The datasets I will be using are publicly available and can be accessed here: <https://ourworldindata.org/covid-deaths> as well as on my GitHub here: <https://github.com/CBW3/Covid-Case-Study>. In particular, the datasets contain information regarding Covid deaths, Covid vaccinations, dates in relation to Covid deaths & vaccinations, and a large variety of demographics data such as populations, countries, etc.

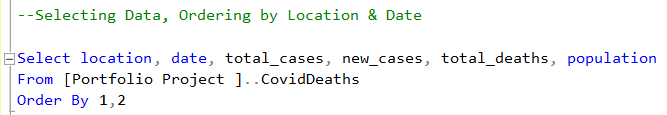
To begin with, I create a new database titled Portfolio Project in Microsoft SQL Server, and I load the datasets into the database using a simple import wizard. Upon doing so, I immediately run a couple queries to make sure that data was imported correctly.



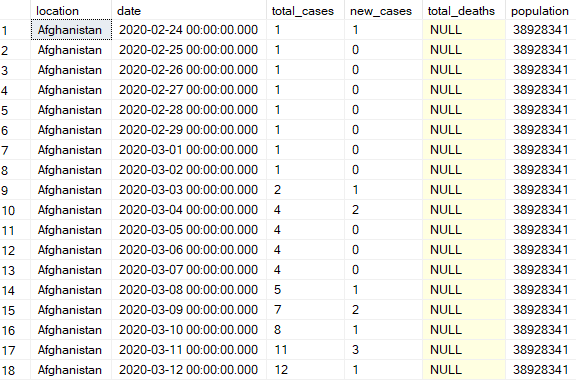
It appears that everything was loaded correctly. I can see the tables clearly, and the information appears to be correct.



Next, I want to particularly focus on the Covid-19 Deaths dataset. I want to perform a simple query to explore disease data, population data, and death data.

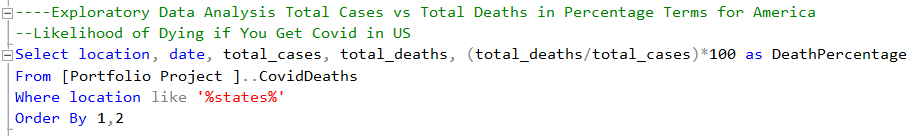


Upon running the query above, I produce the following table:

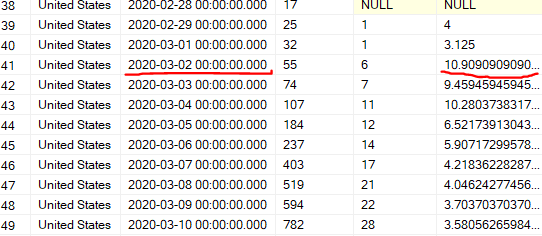


From the table, I can see total Covid-19 cases by date per country. I can also see the number of new cases per day as well as the total population for every individual country.

Now, I want to extrapolate this information to focus on a singular particular country. Since I am a resident of the United States, I am most curious to explore the information as it relates to my home country. Moreover, I actually want to see death rate in the United States because this is the outcome that I am most concerned about. To do so, I actually have to create a new temporary variable titled DeathPercentage using existing information, such as total cases and total deaths.

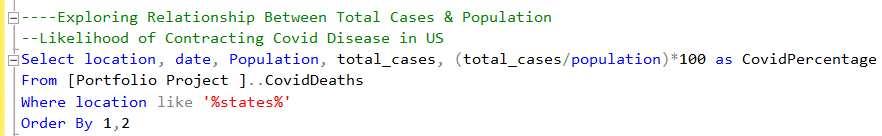


Upon running the query, I produce the following table:

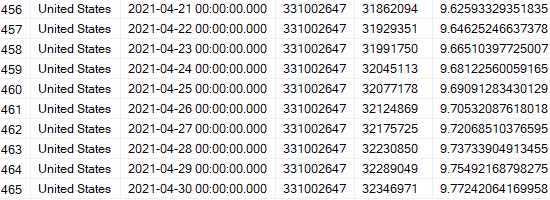


In the table, something in particular stood out to me. On March 2, 2020, the death rate was a whopping 10% of people. Essentially, this information can be interpreted that disease was particular deadly for the affected populace on that day. For some context, this was early on in the inception of the pandemic, so it makes sense that the death rate would be so high. At this point, vaccinations were not readily-available, and there was little to no herd immunity externalities taking place. You can also see that the total cases rapidly increased in a matter of days, increasing from 17 cases to 782 cases.

In relation to the cases rising, I wanted to explore how the relationship between total cases and population within the United States. In particular, I wanted to calculate the percentage of the population affected by the disease.

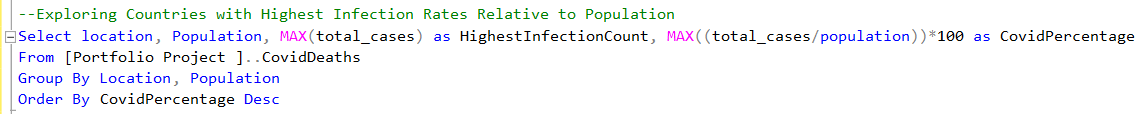


Upon running the query, I produce the following table:

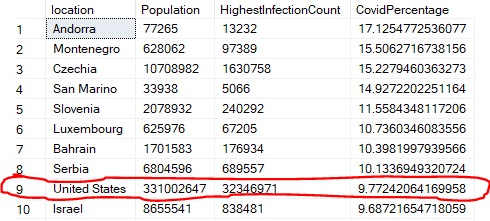


We can see from the table above that the likelihood of contracting the Covid-19 disease in the United States near the heights of the pandemic was shockingly almost 10%!

Previously, I focused the data with respect to the United States. However, now, I would like to view the dataset in a more global lens. This will help me gain some perspective with how the United States was doing in relation to other countries when it comes to the pandemic. Ultimately, I want to explore countries with the highest infection rates relative to the size of their respective populations.

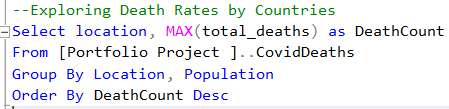


I essentially order the table in descending order as well in order to ensure that the highest infection rates will appear first. Upon running the query, I produce the following table:

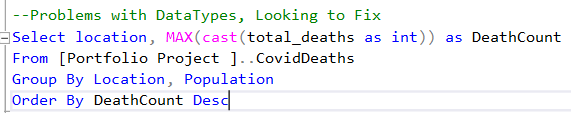


Unfortunately, even on a global scale, the United States is not doing too hot. In fact, the United States is among the top 10 countries with the highest infection rates. In particular, Andorra ranks the highest, with 17% of their total population contracting the disease at one point in time. Not pictured is the 11th ranked country – Sweden. I found this observation particularly interesting because, intuitively, I thought Sweden would be more isolated from the pandemic considering its status as an island nation. However, that appears to not be the case.

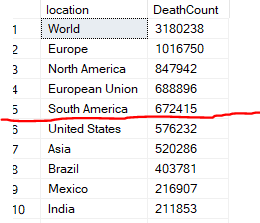
Previously, I explored infection rates. However, now, I want to transition my analysis towards death rates. I suspect that countries with better healthcare, like the United States, may have relatively high infection rates but low death rates. Intuitively, I suspect poor developing countries to have higher death rates compared to wealthy developed countries. To explore this relationship, I run the following query and produce the following table:



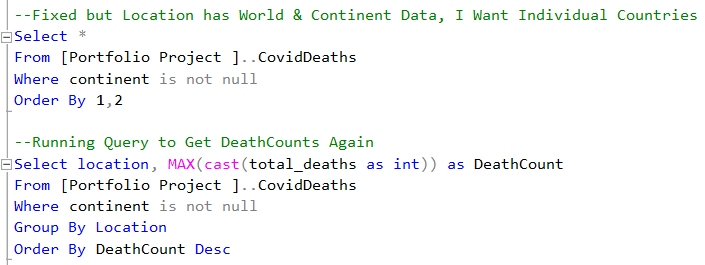
I actually encountered an error when running this query. It was a datatype issue, so I aim to correct that with the next query.



Upon running the newly corrected query, I produce the following table:



So, I immediately notice an issue with this table. I do have countries and their related death counts, but I also have continent related data. I don’t want data with respect to continents, so I aim to correct this issue with the next queries.

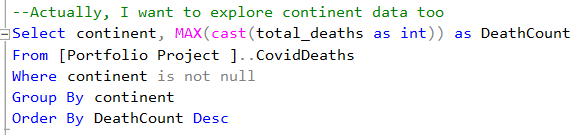


Upon running the above queries, I produce the following table:

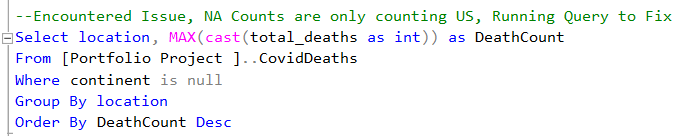


Unfortunately, my intuition appears to be incorrect. In fact, even wealthy well-developed countries like the United States, United Kingdom, France, and Germany all have high death counts. Moreover, the United States is actually the world leader when it comes to Covid-19 deaths. In this paper, I won’t explore how domestic policy effected this outcome; however, naturally, I think domestic policy would have an effect on death outcomes.

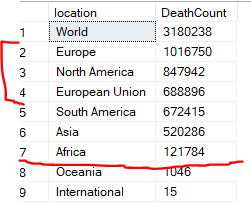
Regardless, I actually want to revisit my analysis in a global context. Earlier, I filtered out the data in such a way that I only got death counts relative to individual countries. However, now, I want to explore death counts by continent. Revisiting the concept of intuition prior, I still suspect that more developed areas will have lower death counts. While it didn’t ring true on a country basis, perhaps, I just need to zoom out a bit. For instance, I think Africa will have the highest death counts by continent due to their large population and poor healthcare programs. And, Europe will have low death counts due to their wealth and strong healthcare programs. To further explore this topic, I run the following query:



I actually encountered the same error as previously, regarding datatype issues. So, I need to restructure my query and correct this problem.

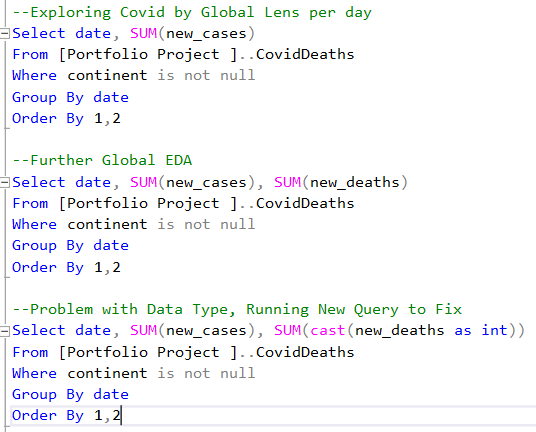


Upon running the query, I produce the following table:

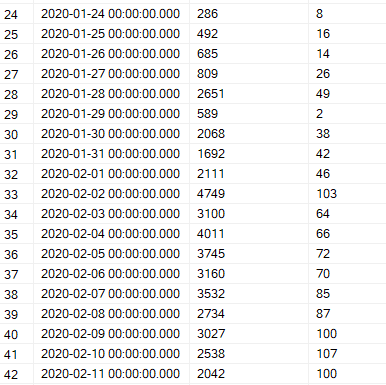


Yeah, this table just confirms, yet again, my intuition was flat out wrong. In fact, wealthier parts of the world with the best healthcare programs actually experienced the highest death counts. And, Africa actually had among the lowest death counts in spite of their large populace. Moreover, Oceania had the lowest death counts. While this makes sense considering their small population size, I suspect this is more directly caused by their status as a collection of islands, isolating themselves from other parts of the world.

Regardless, I still elect to direct my analysis of the pandemic in a more global lens. In particular, I want to explore the relationship between new deaths and new cases per day.

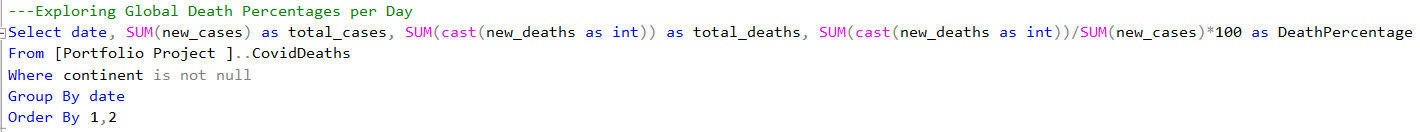


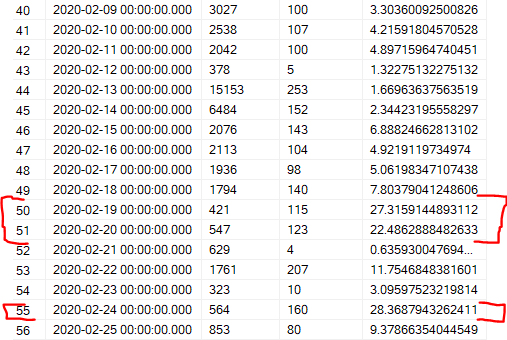
I actually encountered the same datatype error, so I ran a final query to correct that error. Nevertheless, upon running the query, I produce a table showing the counts of new Covid-19 cases and Covid-19 deaths per day on a global scale.



From the table, we can infer that both cases and deaths rose over time, as suspected. The relationship holds true if you continue to scroll down in the table, indicated a strong correlation between deaths and cases (duh!).

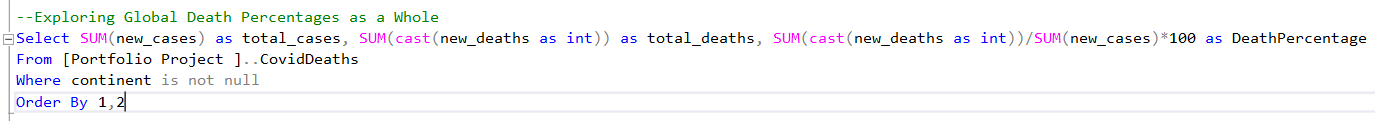
This query really just served as a proof of concept. Now, I want to further extrapolate the information in this table to calculate death percentages.





From the table, I can see the global death rate per day. I can see the date, total cases per day, deaths per day, and the death rate per day. It appears that February 19th, February 20th, and February 24th were particularly deadly days.

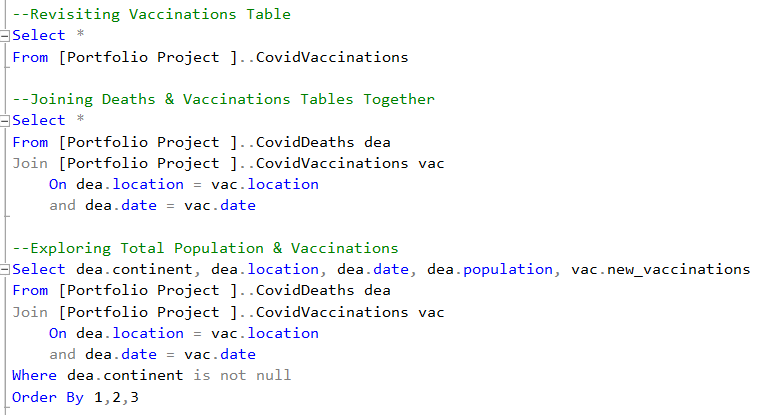
Now, I want to zoom out quite a bit. In particular, I want to see the death rate among all cases of Covid-19 on a global scale.



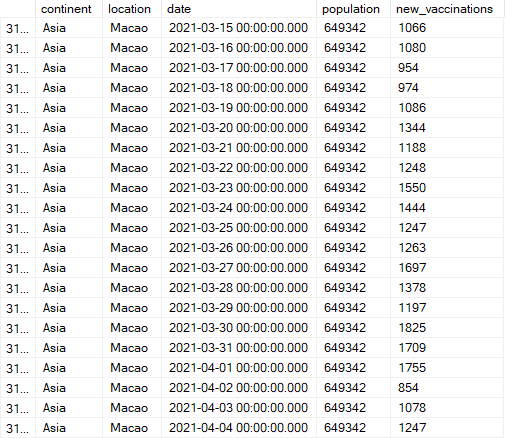


From the table, we can see that there were over 150,000,000 cases of Covid-19, quite a lot. In fact, in excess of 2% of the entire global population contracted the disease. And, I should note, that the data pertains to only confirmed cases!

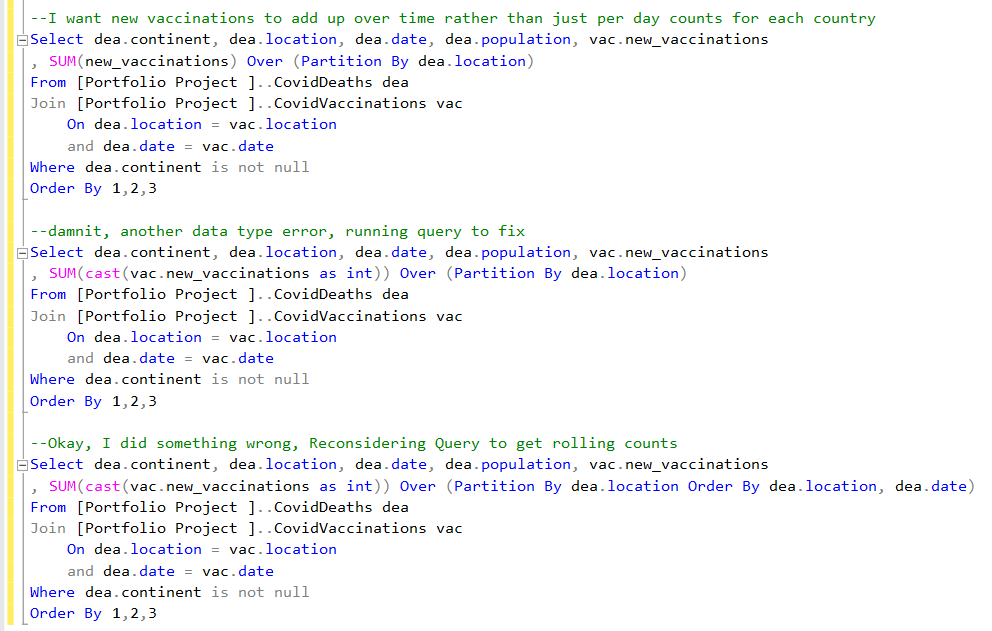
So, most of the analysis above only takes into account death data. However, now, I want to switch gears, and focus my attention more towards the vaccination dataset. In particular, I want to explore the relationship between vaccinations and population over time.



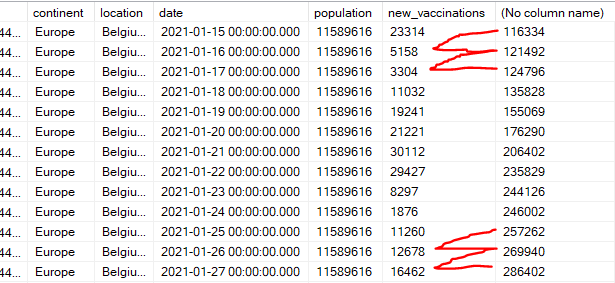
In running the queries above, I join the deaths dataset to the vaccinations dataset. Ultimately, I produce the following table from the final query:



Here, I can see new vaccinations total per day by country. But, what I really want is something akin to a rolling sum of vaccinations over time. This information would be much more useful, so I aim to refine my query to produce what I am looking for.

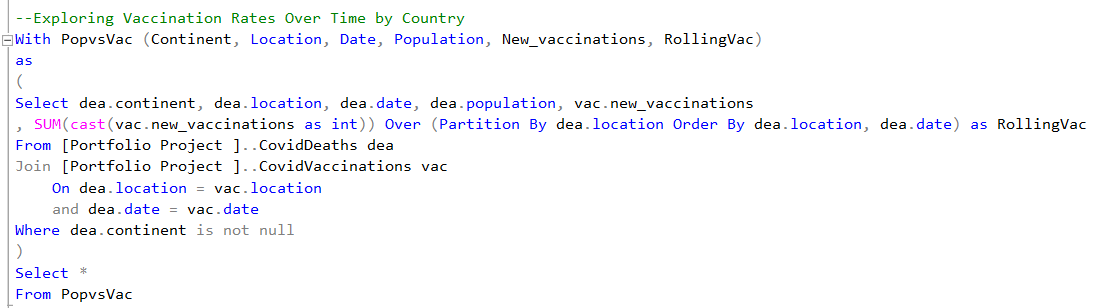


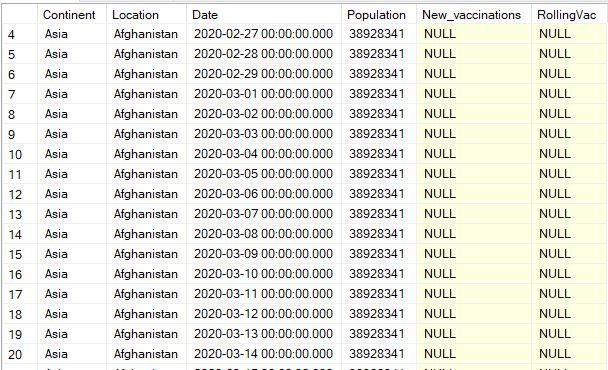
The first two queries have mistakes. I just wanted to include them to document how I corrected & troubleshooted these errors. Nevertheless, upon running the final refined query, I produce the following table:



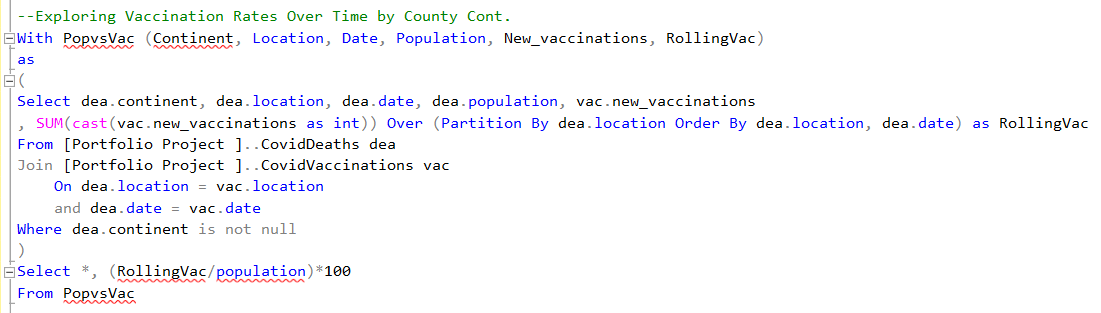
This is the table I was initially envisioning. I have information on the continent, country, date, population, new vaccinations, and a running total on vaccinations. You will notice that, in the first row, 116334 + 5158 = 121492, and so on. This gives me information on how total vaccinations are increasing over time relative to a particular country.

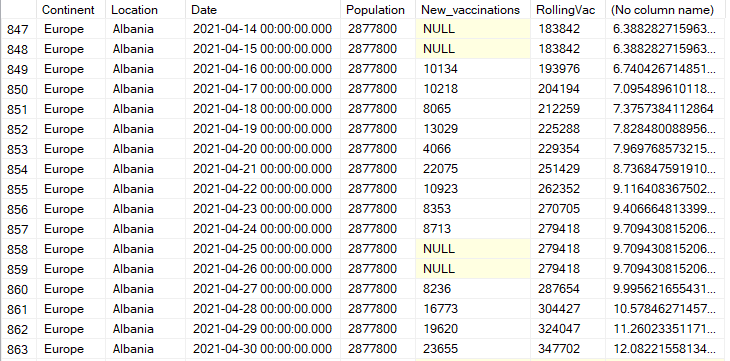
Next, I want to explore the relationship between population and vaccination count over time by country. In order to do so, I need to clean up my query quite a bit. I need to create a new table, so I can perform calculations based upon the new table.





This is the table I want. So, now, I want to transform this table into a new table to perform calculations. I am sure there must be an easier way to perform what I am trying to do; however, given my current competency in SQL, this is the only method I could think of that would work.

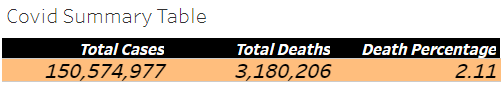




From the table, I can see the percentage of the vaccinated population in a country over time in the final column. As an illustration, given the most recent datapoint, Albania currently has 12% of its population vaccinated against the Covid-19 disease as of April 30, 2021. You will also notice that the percentage increases over time, as expected.

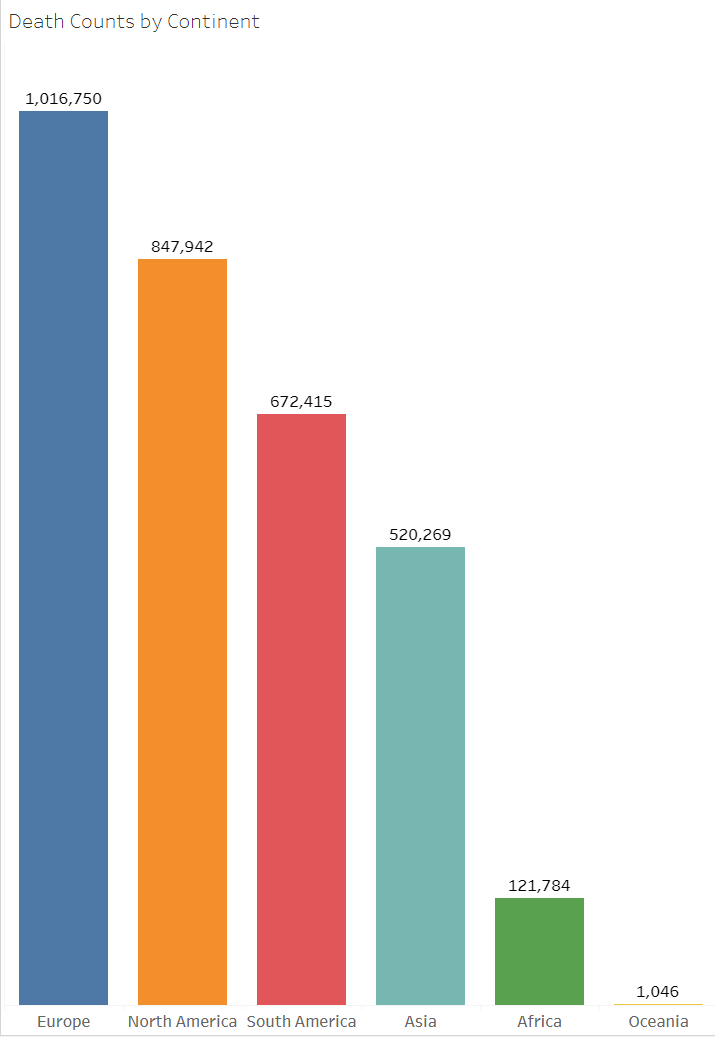
Now, I want to transition to Tableau for some visual analysis. Please, access the link to view data visualizations in a better format and more interactive manner: <https://public.tableau.com/views/Covid-19PandemicVisualAnalysis/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link>.

For my first visualization, I just wanted to produce a very basic summary table.



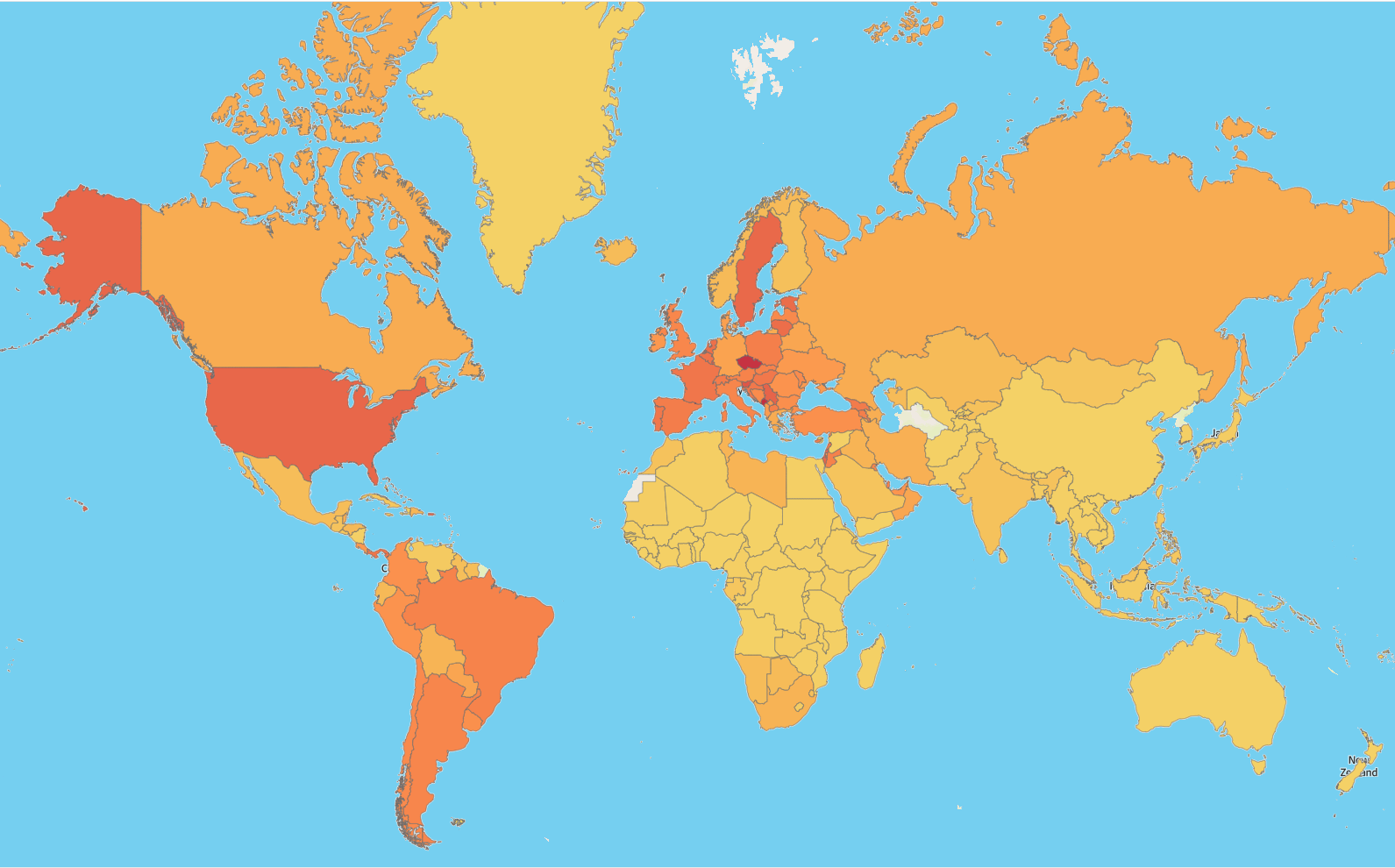
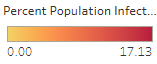
In the table, you can clearly see the total amount of Covid-19 cases from the dataset, the total deaths caused by the virus, and percentage of the afflicted population that die due to the disease.

Next, I wanted to explore the total deaths a bit more in depth, particularly in a geospatial manner. As an illustration, I produced a bar chart highlighting the total amount of deaths per continent.



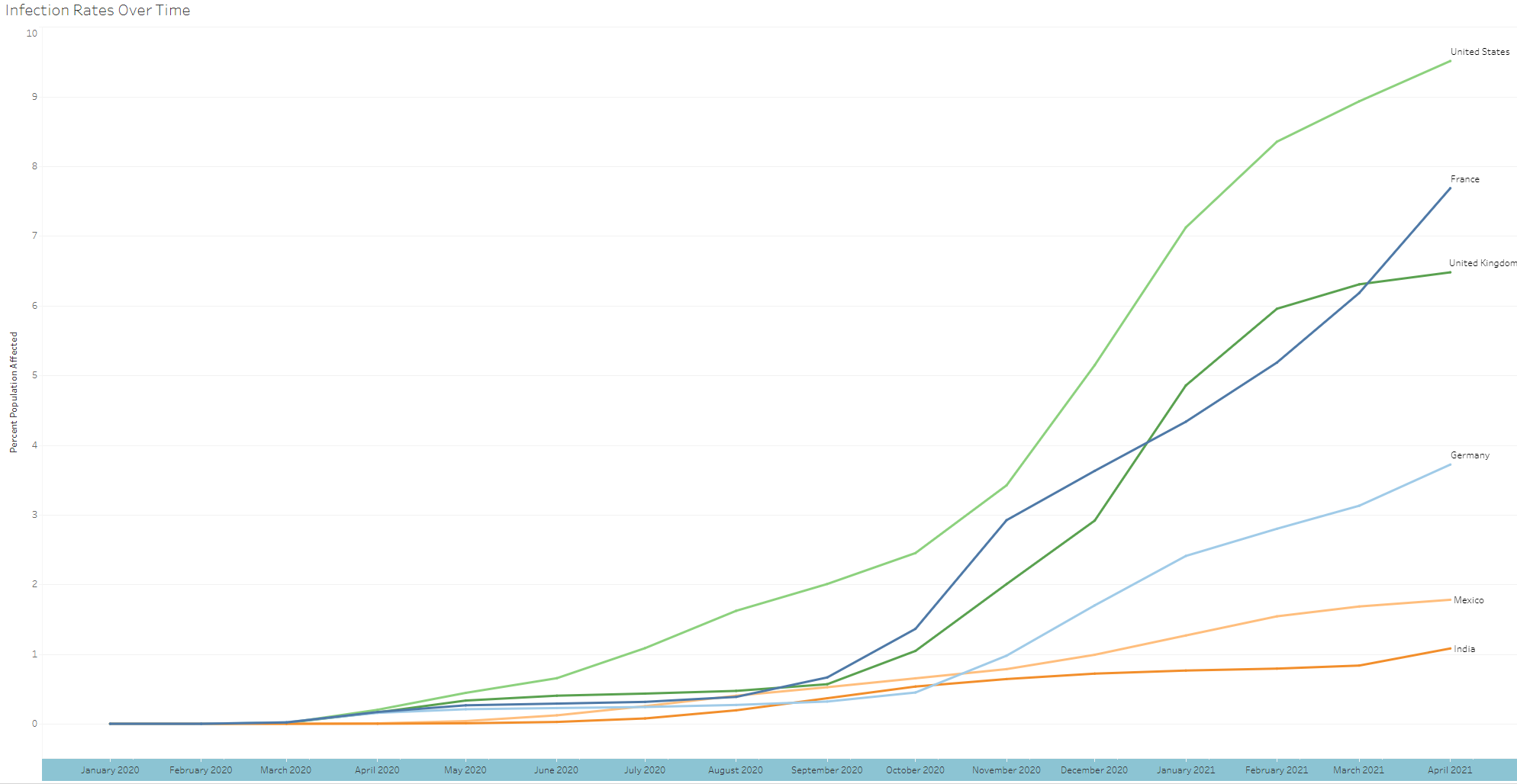
From the visualization, we can clearly see that Europe actually experienced the greatest number of deaths from the virus. Similarly, as we discussed earlier from the SQL exercise, Africa experienced comparatively very little deaths. And, Oceania experienced only ~1k deaths, likely due to their geographic isolation from the virus.

Next, I wanted to explore affliction rates, particularly in a geospatial manner. As an illustration, I produced a density to give us some insight as to where Covid-19 is concentrated.



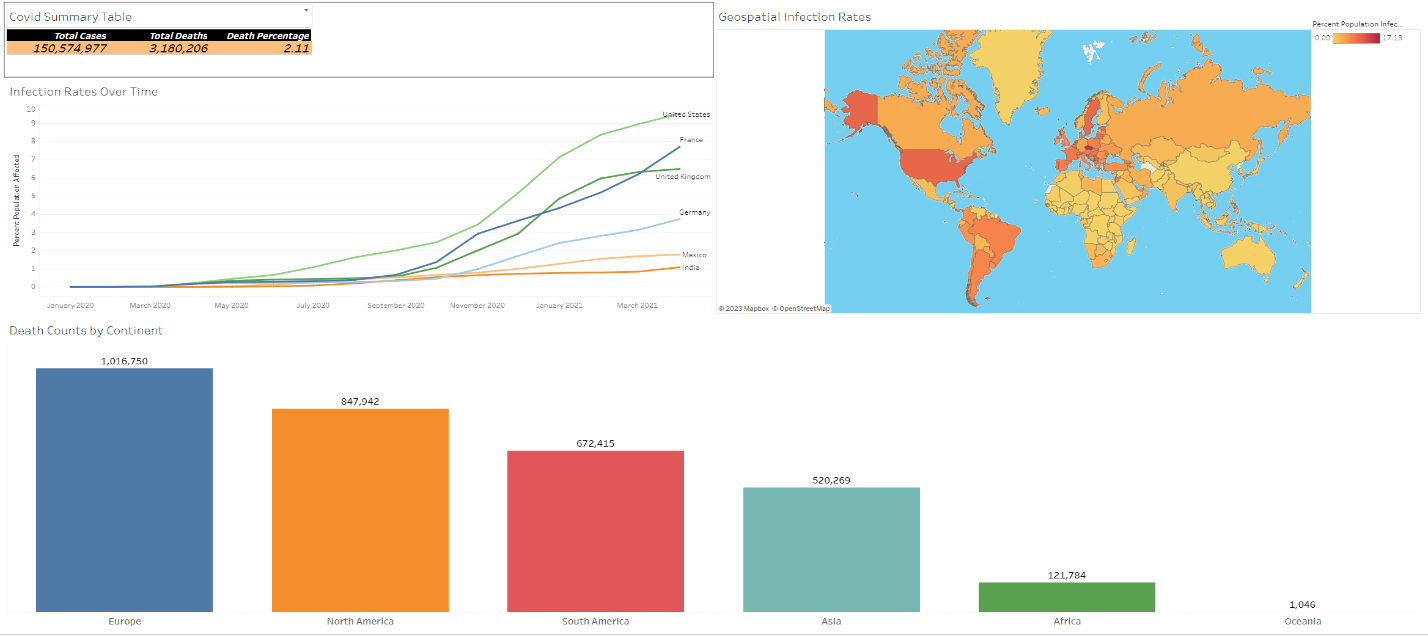
As expected, the United States, in particular, has a very high concentration of Covid-19 instances, represented by the dark red coloring. Similarly, we can see that the island nations, like Australia, Greenland, New Zealand, and Madagascar, were seemingly relatively isolated from the global pandemic, and their respective infection rates reflect that, represented by the light gold coloring.

For my final visualization, I wanted to examine infection rates over time. I picked a handful of countries as a representative sample. Because we are dealing with time-series data for this visualization, I thought a line graph would be most appropriate.



The visualization above display infection rates over time with respect to the United States (light green), France (dark blue), United Kingdom (dark green), Germany (light blue), Mexico (light orange), and India (dark orange). The lines are all rising over time, which is consistent with the nature of Covid-19’s rapid global spread. Likewise, it appears that some of the countries experienced an inflection point of sorts around October 2022 – the infection rates started to rapidly increase beyond this point.

Finally, I produced a dashboard with a collection of all the previous graphs. The dashboard really helps to tell the story of the data in relation to the global pandemic.



In conclusion, in this exercise, I demonstrated my proficiency is using SQL & Tableau. Moreover, I performed a wide variety of functions to conduct exploratory data analysis, and I later utilized Tableau to produce insightful data visualizations in relation to the Covid-19 global pandemic. Please, access the links provided throughout this paper for further information. Thank-you.