




# BIG DATA AND BUSINESS INTELLIGENCE MODULE

## COVID-19 PANDEMIC ANALYSIS

**BY: SALISU ALIYU ABDULLAHI**

**C2224671**

# Table of Contents

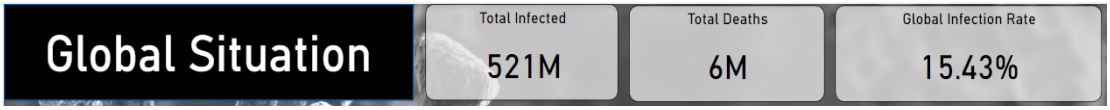
## Executive Summary

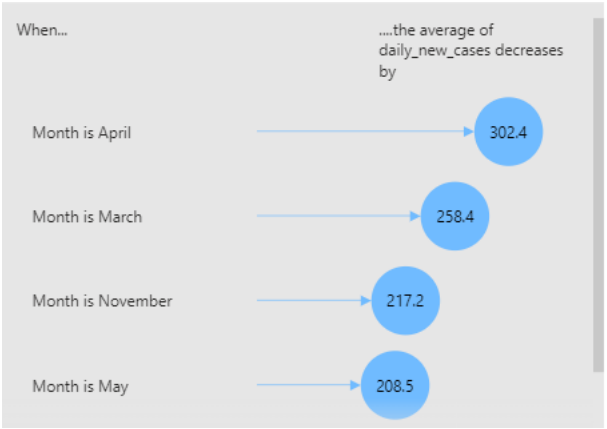
This report examines the coronavirus (COVID-19) pandemic spread since its outbreak in the world. This report investigates confirmed infection and death cases in the world, across the continents, countries and specifically in Nigeria. The purpose of this report is to show how COVID has spread since its outbreak in the world and specifically in Nigeria; and to suggest possible influences that can reduce its spread and aid in the fight to reduce the spread of the pandemic and/or eradicate the disease totally.

### Key Findings

The below are the findings from the business intelligence analysis of the COVID-19 pandemic (as of May 14, 2022,):

- The average COVID-19 infection rate currently stands at 15.43%
- The total number of confirmed infected cases since its outbreak was at five hundred and twenty-one million (521M) with the 2022 being the year with the highest number of infection cases totalling at two hundred and twenty-nine million (229M).
- Over 6 million (6M) people have died from the corona virus pandemic globally with the year 2021 accounting for the largest number of COVID deaths (3.53M) in a year since the pandemic started.
- Europe is the worst hit (194 million) continent by the corona virus with Australia/Oceania being the least hit (8 million).
- Only 19.5 percent of countries in the world have its infection rates below the global infection risk rate.
- Nigeria with a population of over two hundred and fifteen (215) million has an infection rate of 0.12 percent with about five million (5,114,703) covid tests done.
- Since the pandemic, the months of March, April, May, October, and November on average have seen a decline in the daily confirmed infection cases of COVID-19 in Nigeria.



### Recommendations

- There is need for increased attention towards reducing the number of confirmed infected cases in the world as this has been increasing every year.
- Countries across Europe, Asia and North America need to take more stringent measures to reduce the spread of COVID in its regions.
- Lockdown and travel restrictions should be implemented in top countries with high infection risk rates to help reduce the spread of the virus.
- In Nigeria, more tests should be carried out in its population as analysis suggests a high number of the population's infection is not being tested, confirmed, and reflected.
- More research and analysis need to be done to understand why there is a decrease in daily infection in the months of March, April, May, October, and November and how this can be applied to other months of the year.

## Introduction

Since its outbreak at the end of 2019, the corona virus pandemic has become a name on everyone's lips. This is not for a positive reason as the outbreak of the disease put the world on hold for at least month. This disease has affected everything in the world, from businesses to how we interact with each other and the environment. While there has been progress so far in the fight against the disease, there is still much to be done. This project seeks to investigate how the disease has fared so far across various demographics since its outbreak. This is examined based on its impact in the world, across the continents, countries and specifically in Nigeria. Analytical skills such as data pre-processing, modelling, in-depth analysis using AI and visualisations were employed towards this project.

### Data Source and Description

The covid dataset was compiled on woldometers.info by Joseph Assaker and contains covid infection, deaths, and tests information about 225 countries. The organization main goal is to provide easily accessible data on the covid pandemic and assist in the fight to reduce the spread of the infection and possible eradication. Furthermore, the dataset was extracted from Kaggle via the link: https://www.kaggle.com/datasets/josephassaker/covid19-global-dataset.

The dataset is an file containing two sets of data; the first dataset(daily data) is a dataset that shows the daily infections and deaths due to covid across countries in the world while the second dataset (summary data) provides data on the total confirmed cases of covid, the total confirmed deaths due to covid across each country, the severity of these cases, and the number of cases, deaths and tests in relation to each country's population.

Information provided in this dataset was used to measure infection and death rates across various demographic, compare the average infection rate of the world against all the countries in this dataset and examine how the pandemic has fared in Nigeria.

In addition, this data only contains data from February 15th, 2020, to May 14th, 2022, for all countries except China, whose data started from January 22nd, 2020, and Palau whose data started from August 25th, 2021.

Both Datasets are in Comma Separated Values (CSV) format with the daily data set containing 7 columns and 184788 rows while the summary data contains 12 columns and 227 rows. Description of columns contains in the table in each data set (Daily Data and Summary Data) is outlined below.

Table 1: worldometer_coronavirus_daily_data

| Index | Column Name | Description |
|---|---|---|
| 1 | Date | Date information was recorded |
| 2 | Country | Country Location of the information |
| 3 | Cummulative_Total_Cases | Total number of confirmed cases since first case |
| 4 | Daily_New_Cases | Total number of cases confirmed for a particular date |
| 5 | Active Cases | Number of cases that have not recovered or died |
| 6 | Cummulative_total_deaths | Total number of confirmed deaths due to covid since first case |
| 7 | Daily_new_deaths | Total number of confirmed deaths for a particular date |

Table 2: worldometer_coronavirus_summary_data

| Index | Column Name | Description |
|---|---|---|
| 1 | Country | Country Location of the information |
| 2 | Continent | Continent in which the Country Location is in |
| 3 | Total_confirmed | Total number of confirmed cases since first case |
| 4 | Total_deaths | Total number of confirmed deaths due to covid since first case |
| 5 | Total_recovered | Total number of people who have recovered from covid since first case |
| 6 | Active_cases | Number of cases that have not recovered or died |
| 7 | Serious_critical | Number of people left in critical state due to covid infection |
| 8 | Total_cases_per_1m_popupation | Total number of confirmed cases since first case per 1 million of the country's population |
| 9 | Total_deaths_per_1m_popupation | Total number of confirmed deaths due to covid since first case per 1 million of the country's population |
| 10 | Total_tests | Total number of covid tests carried out |
| 11 | Total_tests_per_1m_popupation | Total number of covid tests carried out since first case per 1 million of the country's population |
| 12 | Population | Describes Population of the country |

Below is also a screenshot of the datasets table in excel:

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | country | continent | total_conf | total_deat | total_reco | active_cas | serious_ci | total_case | total_deat | total_tests | total_tests | population |
| 2 | Afghanista | Asia | 179267 | 7690 | 162202 | 9375 | 1124 | 4420 | 190 | 951337 | 23455 | 40560636 |
| 3 | Albania | Europe | 275574 | 3497 | 271826 | 251 | 2 | 95954 | 1218 | 1817530 | 632857 | 2871945 |
| 4 | Algeria | Africa | 265816 | 6875 | 178371 | 80570 | 6 | 5865 | 152 | 230861 | 5093 | 45325517 |
| 5 | Andorra | Europe | 42156 | 153 | 41021 | 982 | 14 | 543983 | 1974 | 249838 | 3223924 | 77495 |
| 6 | Angola | Africa | 99194 | 1900 | 97149 | 145 | | 2853 | 55 | 1499795 | 43136 | 34769277 |
| 7 | Anguilla | North Am | 2984 | 9 | 2916 | 59 | 4 | 195646 | 590 | 51382 | 3368870 | 15252 |
| 8 | Antigua A | North Am | 7721 | 137 | 7511 | 73 | 1 | 77646 | 1378 | 18901 | 190076 | 99439 |
| 9 | Argentina | South Am | 9101319 | 128729 | 8895999 | 76591 | 372 | 197992 | 2800 | 35716069 | 776974 | 45968174 |
| 10 | Armenia | Asia | 422896 | 8623 | 412048 | 2225 | | 142219 | 2900 | 3068217 | 1031834 | 2973558 |
| 11 | Aruba | North Am | 35693 | 213 | 35199 | 281 | | 331689 | 1979 | 177885 | 1653053 | 107610 |
| 12 | Australia | Australia/ | 6593795 | 7794 | 6199822 | 386179 | 129 | 253112 | 299 | 70885598 | 2721042 | 26050899 |
| 13 | Austria | Europe | 4212492 | 18303 | 4135885 | 58304 | 58 | 462804 | 2011 | 1.85E+08 | 20328801 | 9102106 |
| 14 | Azerbaijar | Asia | 792638 | 9709 | 782869 | 60 | | 76885 | 942 | 6838458 | 663324 | 10309383 |
| 15 | Bahamas | North Am | 33871 | 801 | 32488 | 582 | | 84626 | 2001 | 233473 | 583330 | 400242 |
| 16 | Bahrain | Asia | 576997 | 1479 | 569758 | 5760 | 4 | 318491 | 816 | 9775981 | 5396149 | 1811659 |
| 17 | Banglades | Asia | 1953012 | 29127 | 1899419 | 24466 | 1273 | 11643 | 174 | 14051455 | 83767 | 1.68E+08 |
| 18 | Barbados | North Am | 76261 | 433 | 71497 | 4331 | | 264768 | 1503 | 669421 | 2324136 | 288030 |
| 19 | Belarus | Europe | 982867 | 6978 | | | | 104078 | 739 | 13220483 | 1399951 | 9443535 |
| 20 | Belgium | Europe | 4116397 | 31613 | 3941350 | 143434 | 117 | 352324 | 2706 | 33846023 | 2896893 | 11683561 |
| 21 | Belize | North Am | 57896 | 676 | 56761 | 459 | 6 | 140843 | 1644 | 549217 | 1336070 | 411069 |

worldometer_coronavirus_summary

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | date | country | cumulative_total_cases | daily_new_cases | active_cases | cumulative_total_deaths | daily_new_deaths |
| 2 | 15/02/2020 | Afghanistan | 0 | | 0 | 0 | |
| 3 | 16/02/2020 | Afghanistan | 0 | | 0 | 0 | |
| 4 | 17/02/2020 | Afghanistan | 0 | | 0 | 0 | |
| 5 | 18/02/2020 | Afghanistan | 0 | | 0 | 0 | |
| 6 | 19/02/2020 | Afghanistan | 0 | | 0 | 0 | |
| 7 | 20/02/2020 | Afghanistan | 0 | | 0 | 0 | |
| 8 | 21/02/2020 | Afghanistan | 0 | | 0 | 0 | |
| 9 | 22/02/2020 | Afghanistan | 0 | | 0 | 0 | |
| 10 | 23/02/2020 | Afghanistan | 0 | | 0 | 0 | |
| 11 | 24/02/2020 | Afghanistan | 1 | | 1 | 0 | |
| 12 | 25/02/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |
| 13 | 26/02/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |
| 14 | 27/02/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |
| 15 | 28/02/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |
| 16 | 29/02/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |
| 17 | 01/03/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |
| 18 | 02/03/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |
| 19 | 03/03/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |
| 20 | 04/03/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |
| 21 | 05/03/2020 | Afghanistan | 1 | 0 | 1 | 0 | 0 |

worldometer_coronavirus_daily_d

***Figure 1: Screenshot of data in excel***

## BI Requirements\Questions

This project seeks to provide the World Health Organization (WHO) and the Nigerian Centre for Disease Control (NCDC) with information on how corona virus has fared across the world and in Nigeria; and to suggest possible influences that can reduce its spread and aid in the

fight to reduce the spread of the pandemic and/or eradicate the disease totally. The analysis is majorly concerned with the following questions:

- What is the daily infection and death count of the virus in the world across each year since its outbreak and what is the average infection rate globally?
- What is the virus infection spread across the continents, how many deaths has it led to and what are the worst hit and least hit continents by the virus?
- What percentage of countries are below and above the global infection rate average and what are the top 10 countries with high risk of infection?
- To show the daily count distribution of infection and deaths due to the virus in Nigeria since outbreak
- To identify the periods which there is a decline in the number of daily infections in Nigeria

## Finding based on analysis and evaluation

The Global situation dashboard below visualises the result of analysis carried out to answer question one. The visualisation contains a drill down timeline zoom chart of Confirmed infected cases and confirmed deaths. These charts show the breakdown of covid infected cases by year (2020 to date). Furthermore, these zoom charts allow drill down into any year of choice should you need further analysis of covid cases based on the months of the year and much further into each day of any month. This chart was used because it is an interactive chart as it provides specific details into any selected period as shown below. The drill down timeline zoom chart shows increase in daily infected cases yearly with 2022 having the highest cases of infection. On the drill down timeline showing death cases, 2021 had the highest death cases when compared to the other years. The dashboard also contains visuals of three cards which show the total number of confirmed infected cases, the total number of deaths due to covid and the average infection rate which from the card stands at 15.43 percent globally.

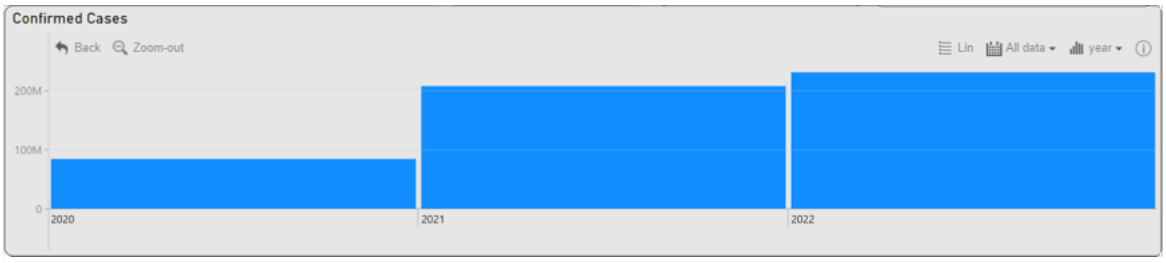



***Figure 2: Distribution of confirmed infection cases by year***

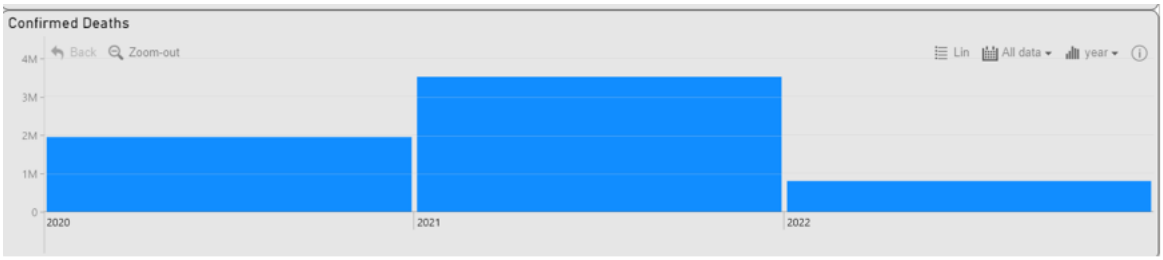



***Figure 3: Distribution of confirmed covid deaths by year***


*Figure 4: Overview of Total Infection, deaths, and Infection Rate*

The Continental Situation dashboard visualises the result of analysis carried out to answer question two. The visualisation contains an aster plot of infection cases distribution across continents in the world. From the plot, we can infer that Europe was the worst hit continent by the corona virus with over 194 million infection cases over the period and Australia/Oceania continent the least hit by the virus with about 8 million infection cases over the same period. Another visualisation in this dashboard (Clustered bar chart) shows the confirmed deaths due to covid across continents in the world. This chart shows that Europe witnessed the highest number of deaths due to the virus with 1.83 million people dying from the pandemic. Also, Australia/Oceania witnessed the lowest number of deaths due to covid with its number at a low of less than a million (about ten thousand). A slicer was added to the dashboard to aid easy navigation and highlight any particular year of choice in the period. A multi-row card was also visualised to aid in highlighting the list countries that fall under each continent under this report.

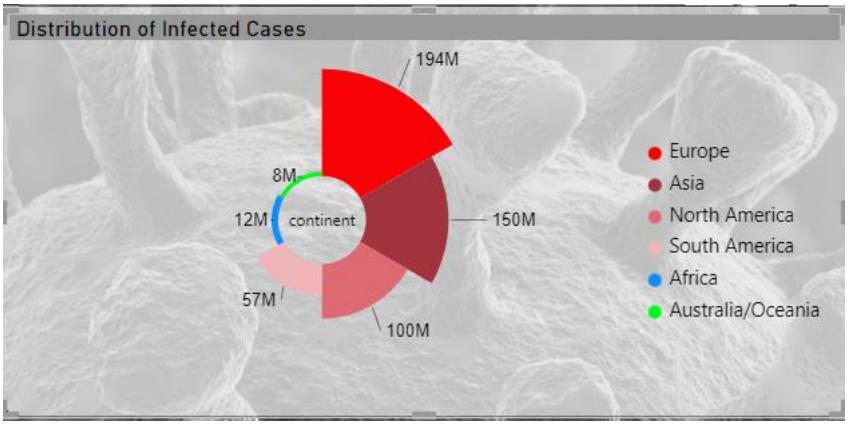



*Figure 5: Distribution of infection cases across Continents*

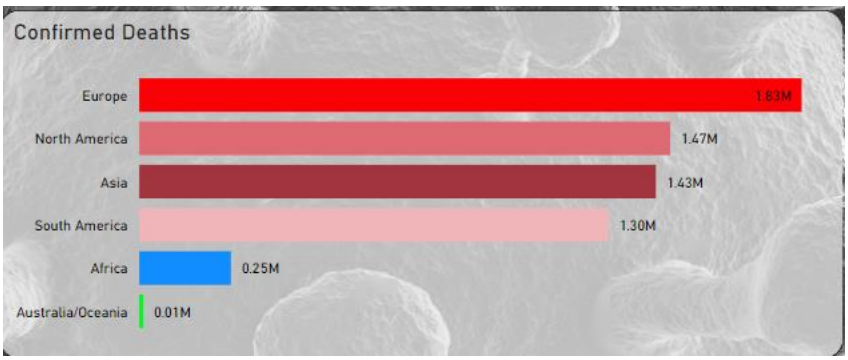



*Figure 6: Distribution of covid deaths across Continents*

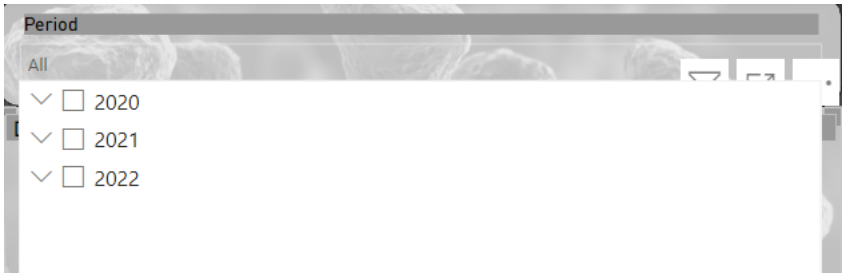



*Figure 7: Period Slicer*

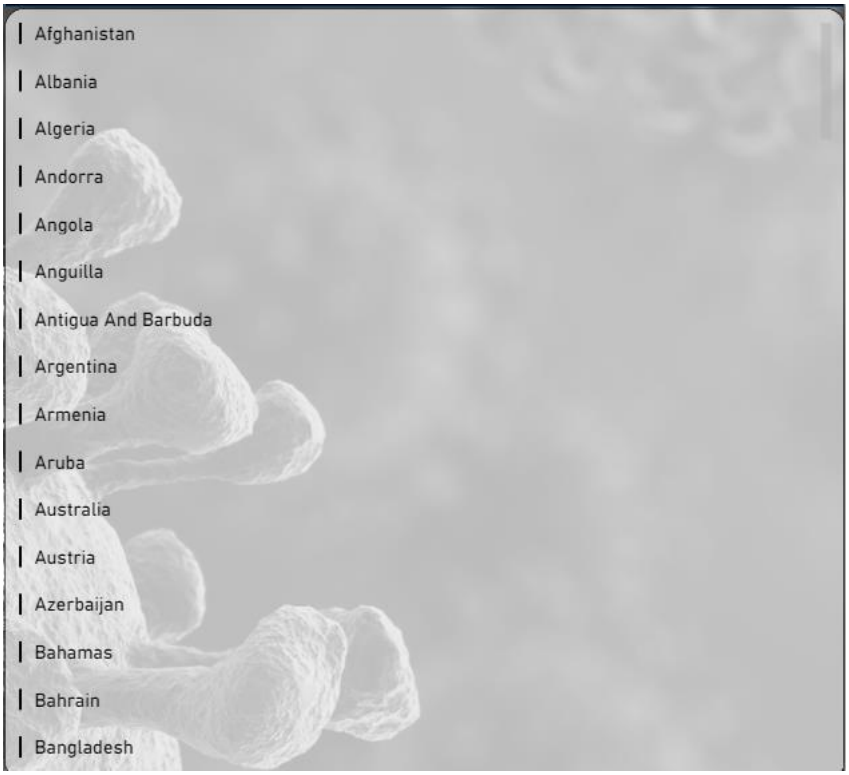



*Figure 8: Card showing List of Countries*

The Country Situation dashboard was created to visualise the result of analysis carried out on the data to answer question three. The visualisation contains a table, a donut chart, and a funnel chart. The table was used to show vital information for each country with regards to the corona virus. The table contains columns which houses information such as country, total infected people, total deaths, infection rate of the country, the mortality rate of covid in the country, the recovery rate of people infected from covid. The table also comes with an identification colour to show countries whose infection rate are above the global average infection rate (red) and countries who are below as well (green). The donut chart visualisation shows the percentage of countries spread according to their infection risk category. From the visualisation, we can infer that only about 19.5 percent of countries in the world are in the low-risk category (between 0-15%) with the largest risk category being countries that have an infection rate between 30 – 50 percent (46.59). The funnel chart visualises the top 10 countries with the highest infection rate of covid. From the chart, Faroe Island has the highest rate of infection with a 70.43% infection rate and Saint Barthelemy at the 10$^{th}$ position with an infection rate of 46.42%.

| country | Total_Infected | total_deaths | Infection_rate | mortality_rate | Recove |
|---|---|---|---|---|---|
| Afghanistan | 179290 | 7685 | 0.44% | 0.02% | |
| Albania | 275559 | 3496 | 9.60% | 0.12% | |
| Algeria | 265815 | 6875 | 0.59% | 0.02% | |
| Andorra | 42155 | 153 | 54.40% | 0.20% | |
| Angola | 99187 | 1898 | 0.29% | 0.01% | |
| Anguilla | 2644 | 8 | 19.56% | 0.06% | |
| Antigua And Barbuda | 7702 | 136 | 7.76% | 0.14% | |
| Argentina | 9101310 | 128728 | 19.80% | 0.28% | |
| Armenia | 422895 | 8623 | 14.22% | 0.29% | |
| Aruba | 35600 | 212 | 33.17% | 0.20% | |
| Australia | 6551609 | 7773 | 25.31% | 0.03% | |
| Austria | 4209155 | 18303 | 46.28% | 0.20% | |
| Azerbaijan | 792637 | 9709 | 7.69% | 0.09% | |
| Bahamas | 33850 | 800 | 8.46% | 0.20% | |
| Bahrain | 576992 | 1479 | 31.85% | 0.08% | |
| Bangladesh | 1952965 | 29126 | 1.16% | 0.02% | |
| Barbados | 76205 | 432 | 26.48% | 0.15% | |
| Belarus | 982866 | 6978 | 10.41% | 0.07% | |
| Belgium | 4116396 | 31613 | 35.23% | 0.27% | |
| Belize | 57889 | 675 | 14.08% | 0.16% | |
| Benin | 26926 | 162 | 0.21% | 0.00% | |
| Bermuda | 14096 | 133 | 22.85% | 0.22% | |
| **Total** | **520780007** | **6287453** | | | |

*Figure 9: Table showing Countries with their Infection rates*

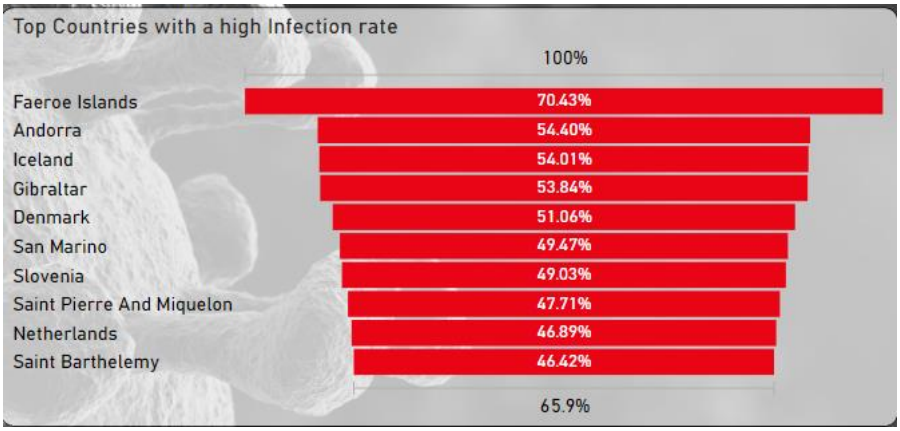



*Figure 10: Top Countries with the highest infection rate*

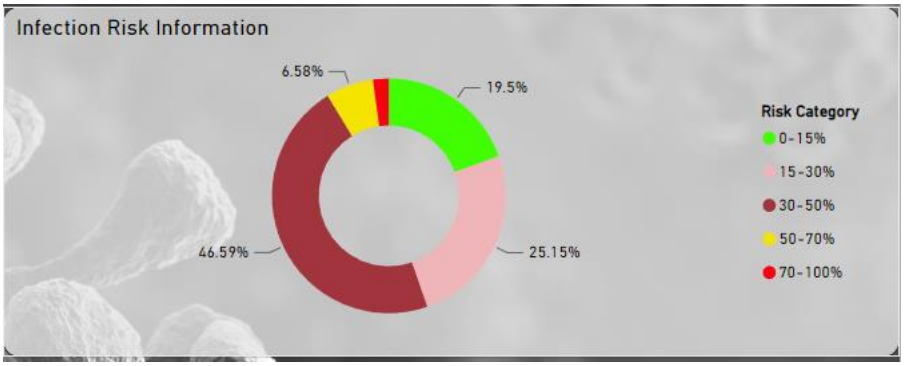



*Figure 11: Distribution of Countries across Infection risk Categories*

The Country-specific situation dashboard was created to visualise results of analysis carried out to daily count distribution of infection and deaths due to the virus in Nigeria since outbreak and identify the periods in which there were decline in the infection number in Nigeria. The visualisation contains a multi row card which gives shows the population of Nigeria (about 215 million), the total number of confirmed infection cases (over 250,000), the total number of deaths due to covid (3,143), the total covid tests done (about 5 million) and the infection rate of covid in Nigeria which stands at 0.12%. There are also two area charts which shows the daily reported infection cases and the daily confirmed covid deaths across the country respectively since the outbreak of corona virus. From the area charts, we can infer that the highest daily infection number in Nigeria was between December 2021 and January 2022. The dashboard also contains a key influencer chart which shows the periods which the average daily number of confirmed infection cases in Nigeria has reduced. From the visualisation, the daily infection cases are likely to reduce during the months of March, April, May, October, and November. Lastly, for easy navigation, there is a Timeline 2.4.0 slicer which is of advantage should you want information specific to a range of period since the outbreak of Covid in Nigeria.

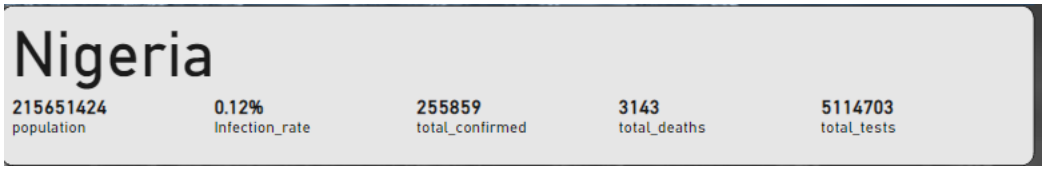



*Figure 12: Overview of Covid infection in Nigeria*

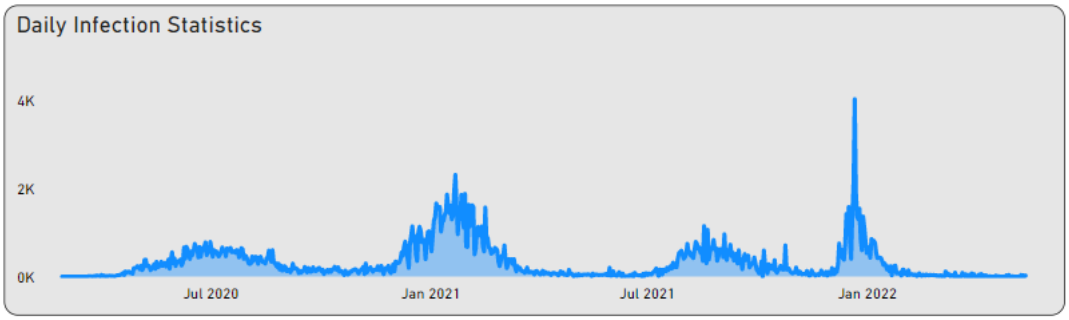



*Figure 13: Daily Covid Infection in Nigeria since outbreak*

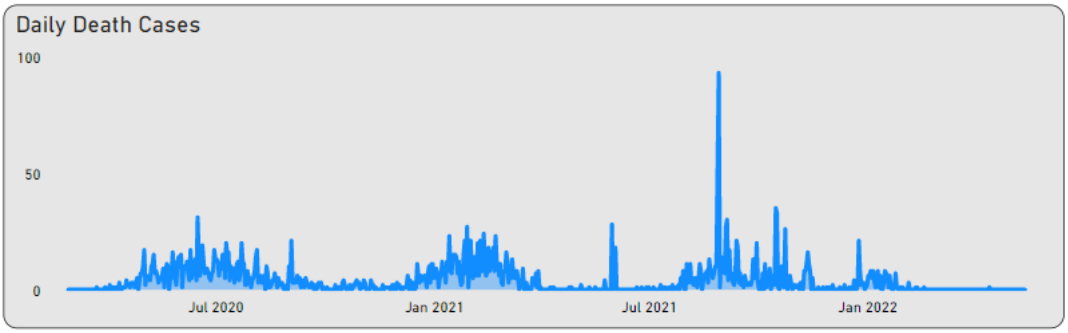



*Figure 14: Daily Covid deaths in Nigeria since outbreak*

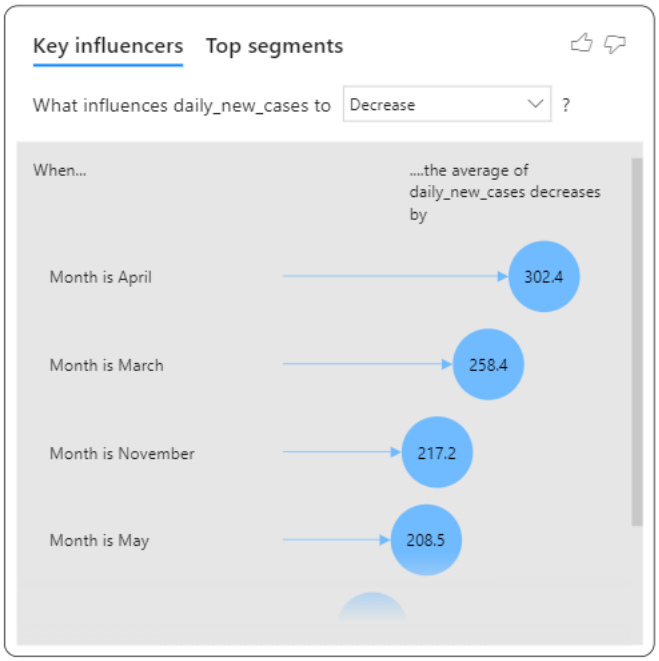



***Figure 15: Period when there is Decline in Daily infection***

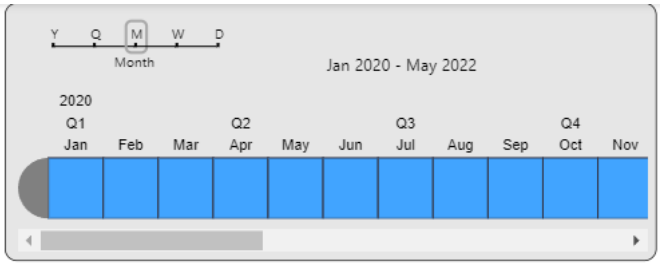



***Figure 16: Timeline Slicer***

The Information Summary dashboard was created to help users gain insights to the graphs and charts used in other dashboards. This dashboard was created using the smart narration visualisation to get information on every dashboard. From the visual, the longest period of rise in infection cases was from January 2020 to July 2020 with the longest drop in deaths being from August 2021 to May 2022. The longest period of drop in infection cases in Nigeria was from January 9, 2022, lasting 125 days.

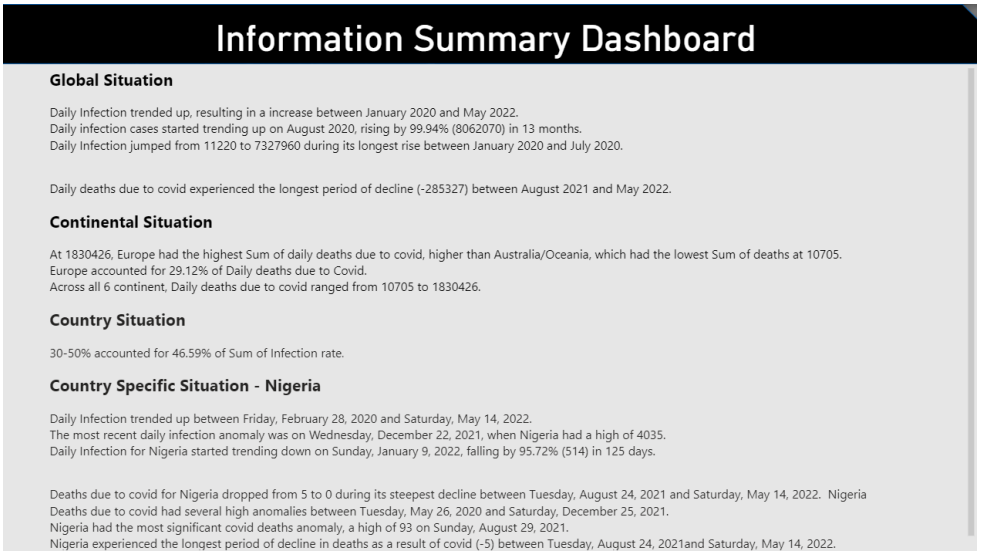



***Figure 17: Smart Narration of Dashboards***

## Recommendations

Based on the findings from the visualisations created. Below are my recommendations.

- While good work has been done so far to reduce the number of deaths caused by covid, there is more work to be done to reduce the number of daily infections of covid as it has been increasing with every year. This would also have a positive effect on the global infection rate.
- In Europe, Asia and North America, there is need to intensify measures to reduce the number of covid infections and consequently deaths as a result as well as the statistics in these continents are higher compared to the remaining continents.
- More efforts need to be added towards improving the distribution percentage of countries below 15% infection risk category by reducing the spread across the other infection risk especially countries between 30-50 percent infection rate.
- Strict measures need to be put in place to effectively reduce the infection rate in the top 10 countries with the highest infection risk.
- In Nigeria, more tests should be carried out per million in its population as analysis suggests a high number of the population's infection is not being confirmed and reflected.
- More research and analysis need to be done to understand why there is a decrease in daily infection in the months of March, April, May, October, and November and how this can be applied to other months of the year.

## Conclusions

I use Excel for analysis and reports before this module and only became aware of power BI through this module. However, I was fascinated by the extensive amount of analysis I can do with Power BI and was mostly impressed by features such as data pre-processing which makes it easier for me to streamline data to be precise and accurate; data modelling which allows me to group my data in smaller sets and still have them connected; the extensive library of visuals and the integration of AI and machine learning tools which are particularly

interesting to use. Finally, the issue I had creating this report mostly had to deal with the delay and lagging at random intervals by power BI when processing the data however I have not been able to determine if this is based on one's personal computer specs. Overall, power BI is a tool with many advantages especially with regards to data analysis.

# Appendix: BI Design

## Data Pre-Processing or Data Cleansing

- Data Loading

The pre-processing stage of the COVID dataset begins with importing of the dataset into the Microsoft power BI application. This is done using the 'Get Data' menu found on the '**Home'** tab of the application to select the file and import. The pictorial steps are shown below

1. Click on **'Get Data'**

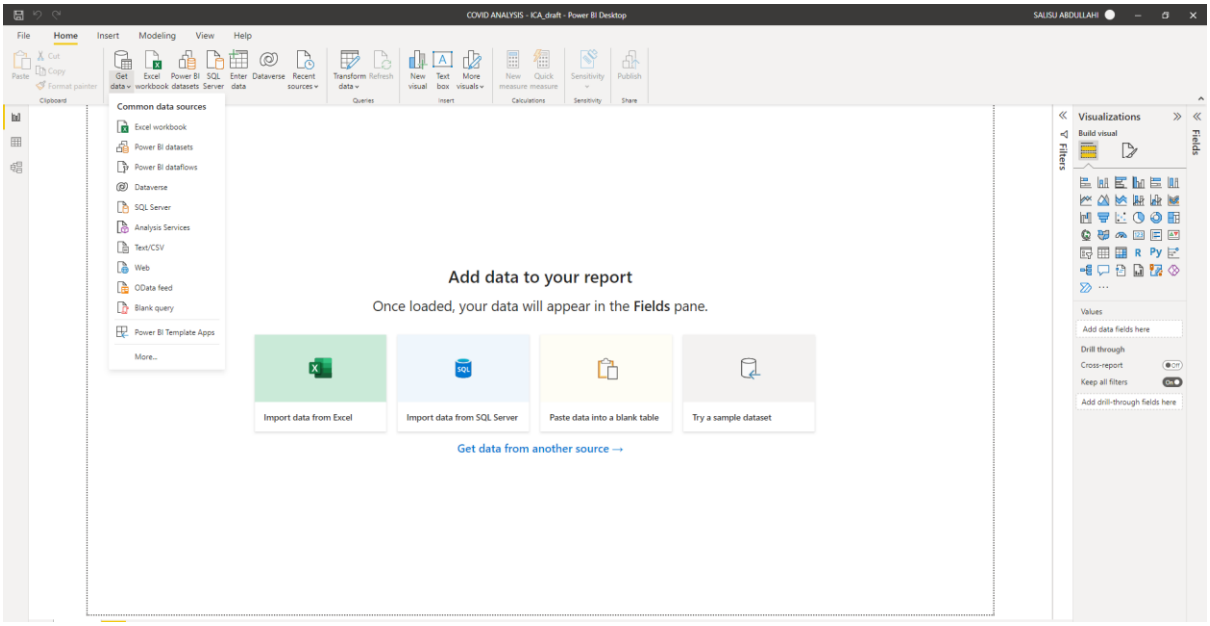



***Figure 1.0: 'Get Data' process***

2. Select appropriate data format. **'Text/CSV'** was selected as dataset is in **'Microsoft Excel Comma Separated values'** format.

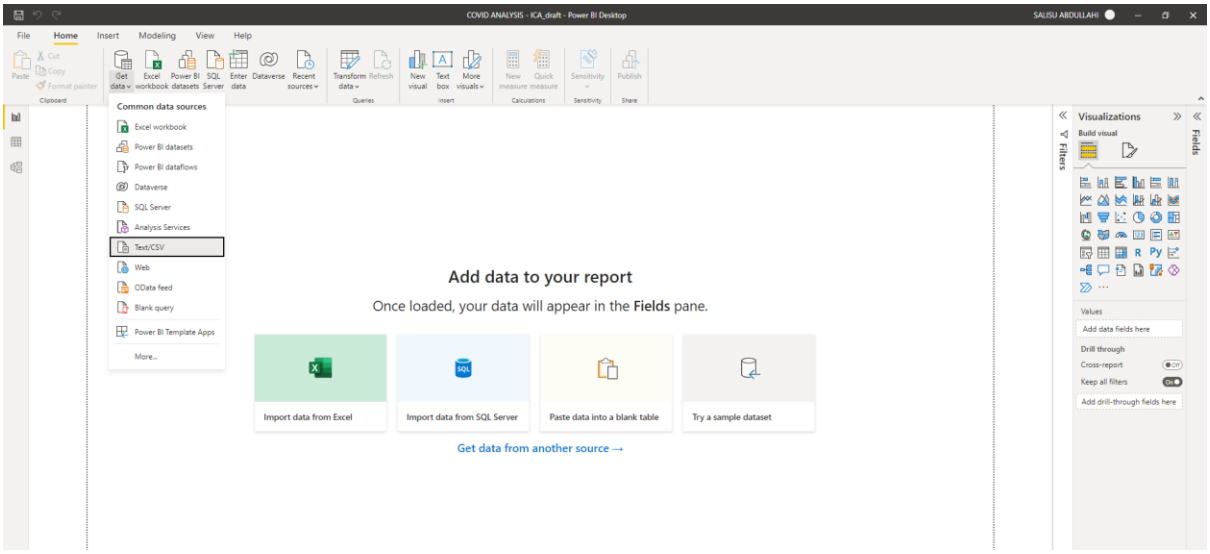



***Figure 1.1: 'Text/CSV' selection***

3. Select dataset files from the directory it was saved on the system and click on open.

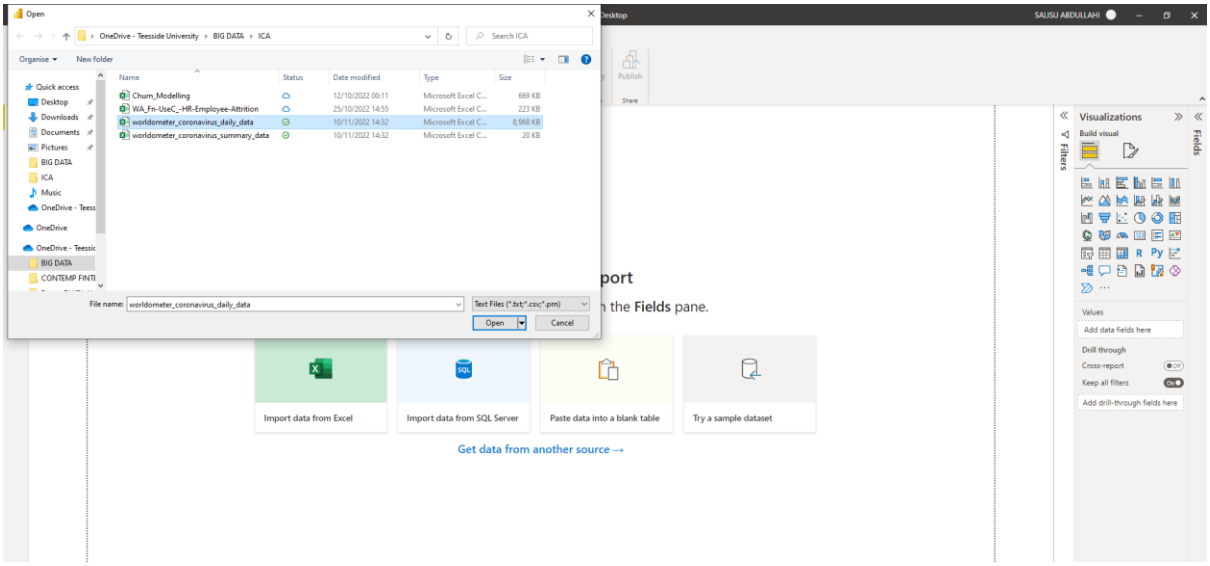



***Figure 1.2: File selection***

4. After selecting the file, a dialogue box opens showing a preview of the dataset selected. This also shows the file origin, the delimiter which the file is separated by (commas as file is a CSV file) and the basis on which the data type was detected.

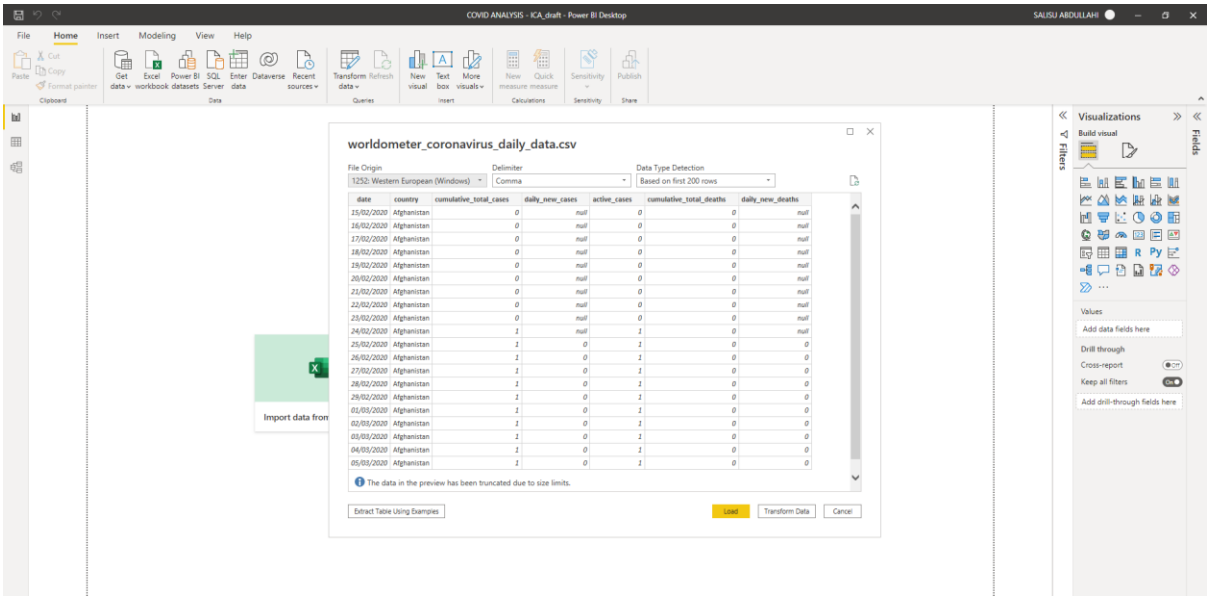



***Figure 1.3.0: Dataset Preview***

5. Upon review of the dialogue box to further confirm that right dataset was selected, the load option was selected.

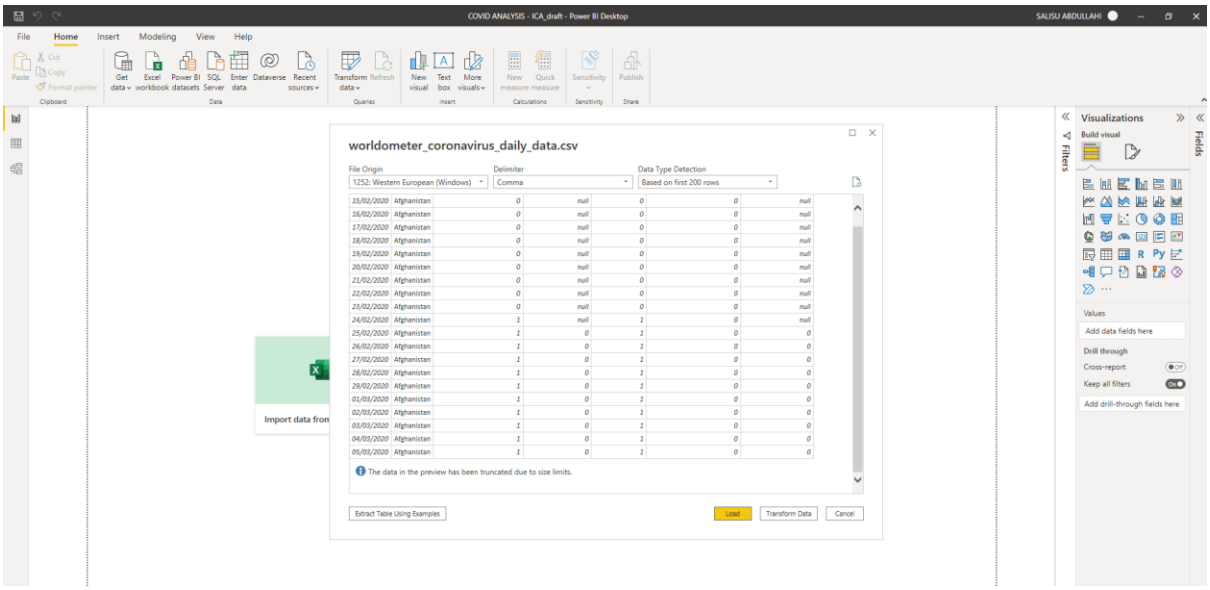



***Figure 1.3.1: Dataset Preview***

6. Steps 2 to 5 was repeated to select the second dataset file to be used after which we have below page as shown. The imported datasets can be found under the fields menu as shown below.

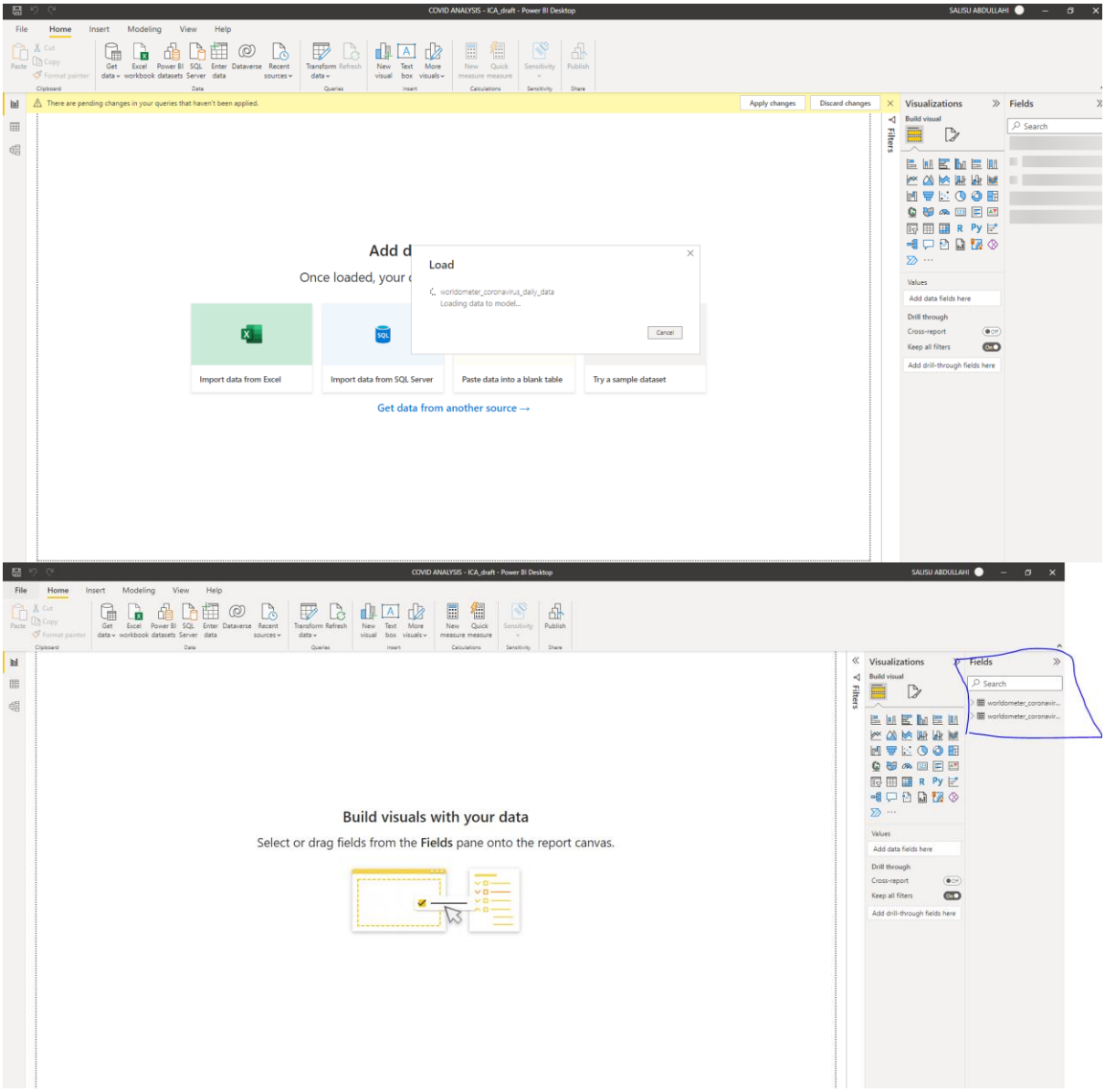



***Figure 1.4: Data loaded into Power BI***

- Data Cleaning

After importation of the dataset, the dataset was then cleaned. This process allows for me to identify errors and fix them as well as remove duplicated data and columns that are not important to the analysis to be done.

Data Cleaning underwent the following procedures.

1. The **Transform data** icon was selected from the query group to open the power query editor

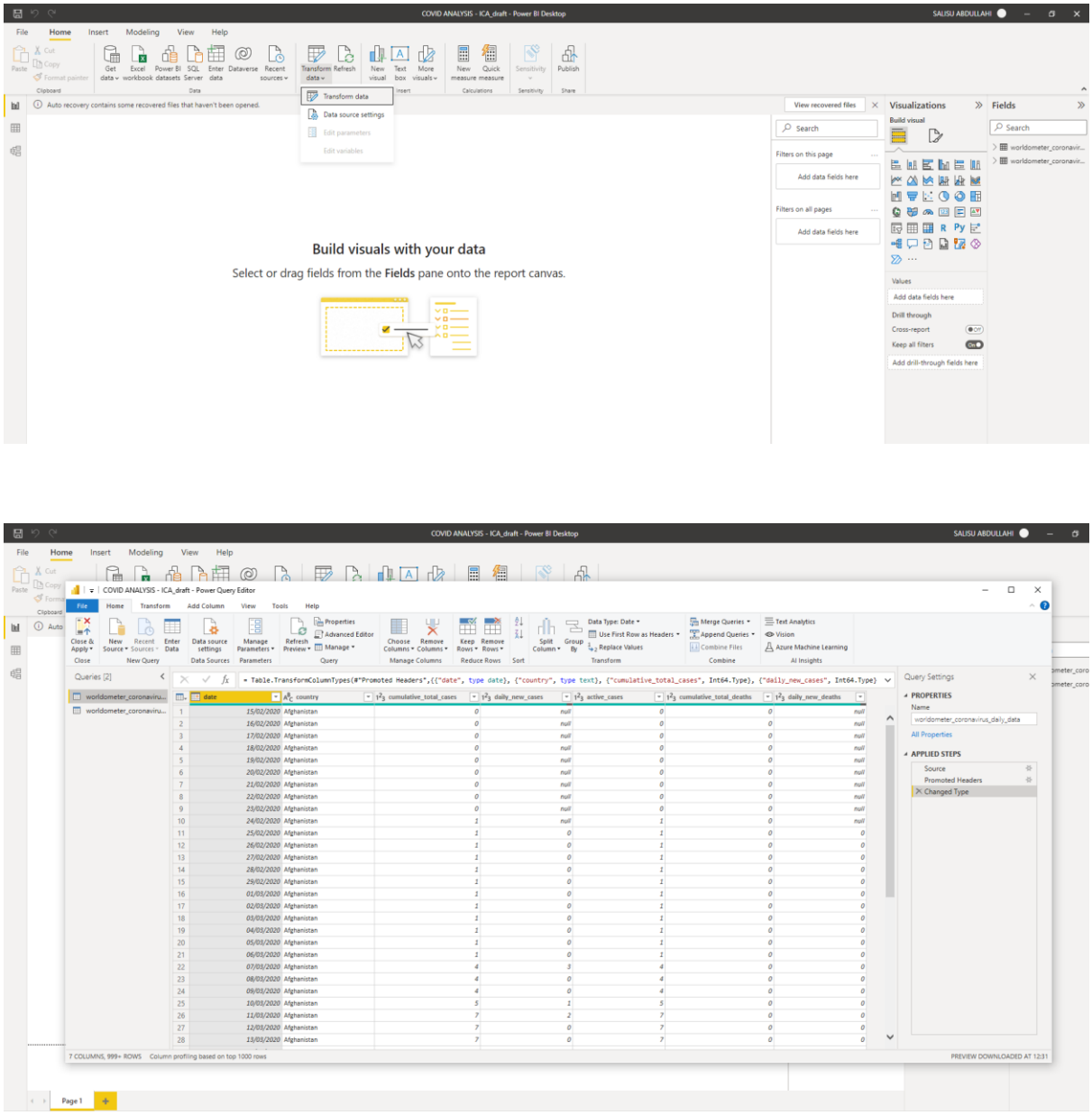



***Figure 2.0: 'Transform Data' selection***

2. The next step was to identify blank, null or error data from our columns. To do this, I first used the **'Column Quality'** under the View Menu bar to identify the percentage of data under each column which can be identified as blank, null or error. The output of this procedure is shown below. The aim of this procedure was to decide the best way to deal with blank, null, and erroneous data. If blank, null, or erroneous data in a column was less than 10 percent of the data, data will be dealt by elimination; if blank data is 10 percent or more than 10 percent, data will be dealt by identification.

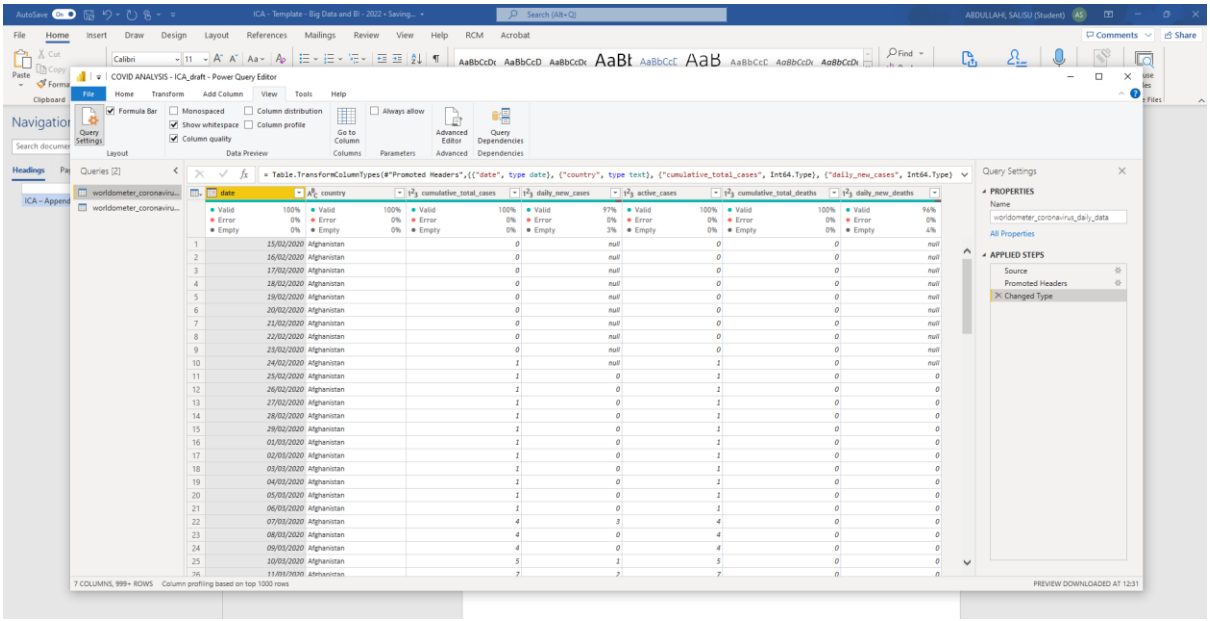
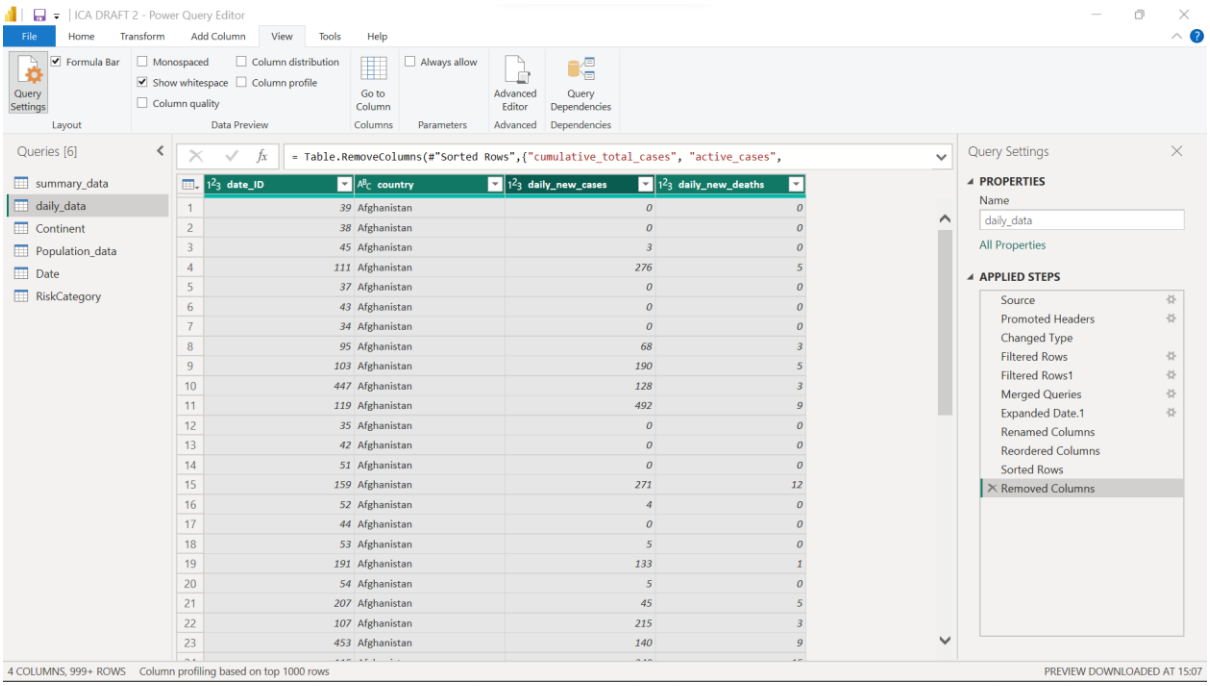


*Figure 2.1: 'Column Quality' selection*

3. Based on the column quality, I decided to eliminate the blank data in each column of my data. To do this, I first selected all columns, then clicked on the dropdown icon beside the **daily_new_cases** column and selected the **'remove empty'** filter option. This removed the null values in my data and gave resulting column quality as shown below. Negative figures in the column were filtered out too. This was done by repeating same step but instead of selecting 'remove empty', I unticked the negative values off the list.



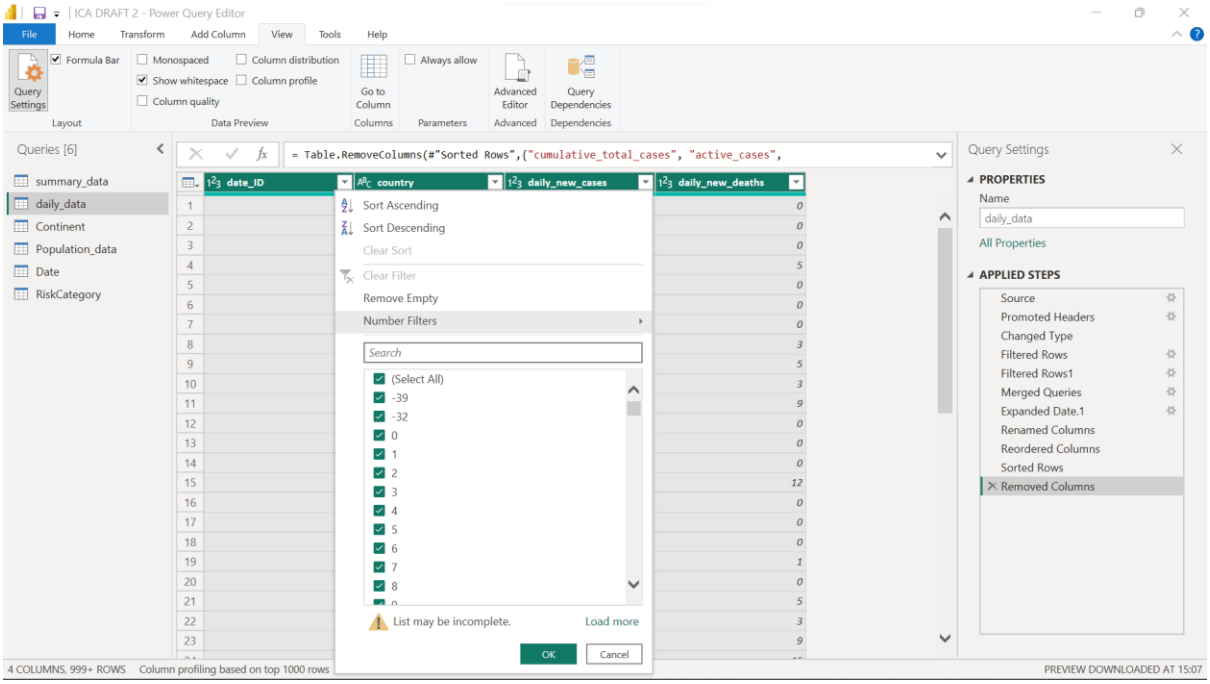



***Figure 2.2: Removal of null values***

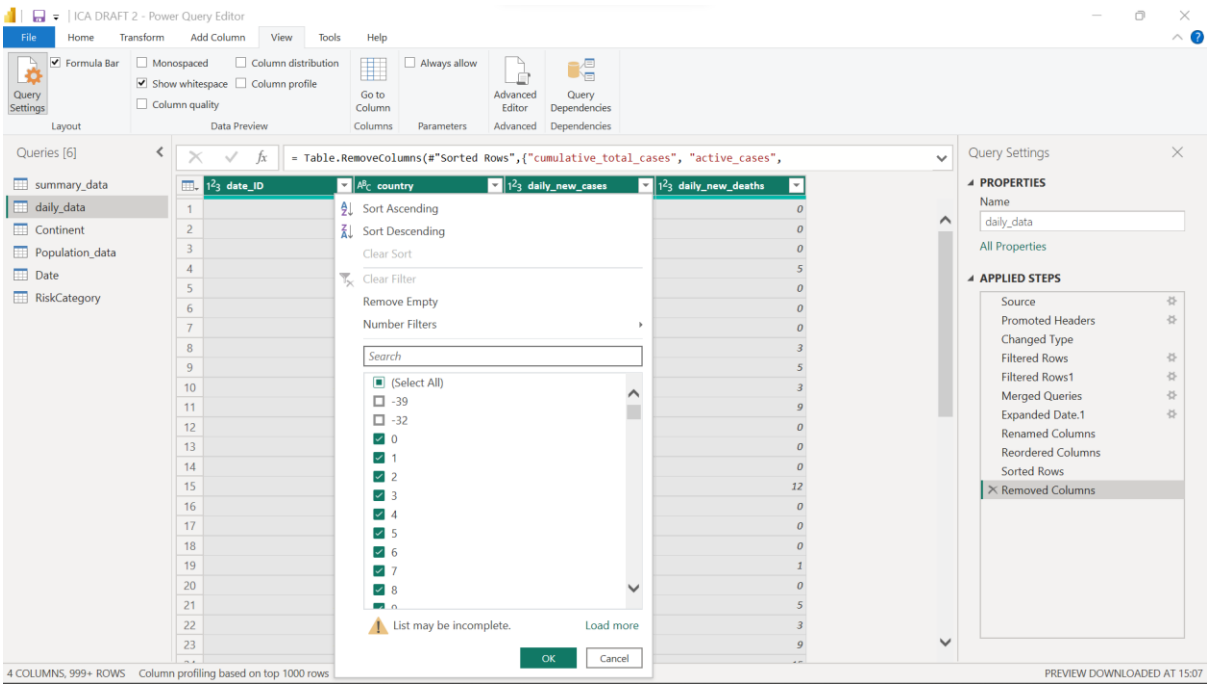



***Figure 2.3: Removal of negative values***

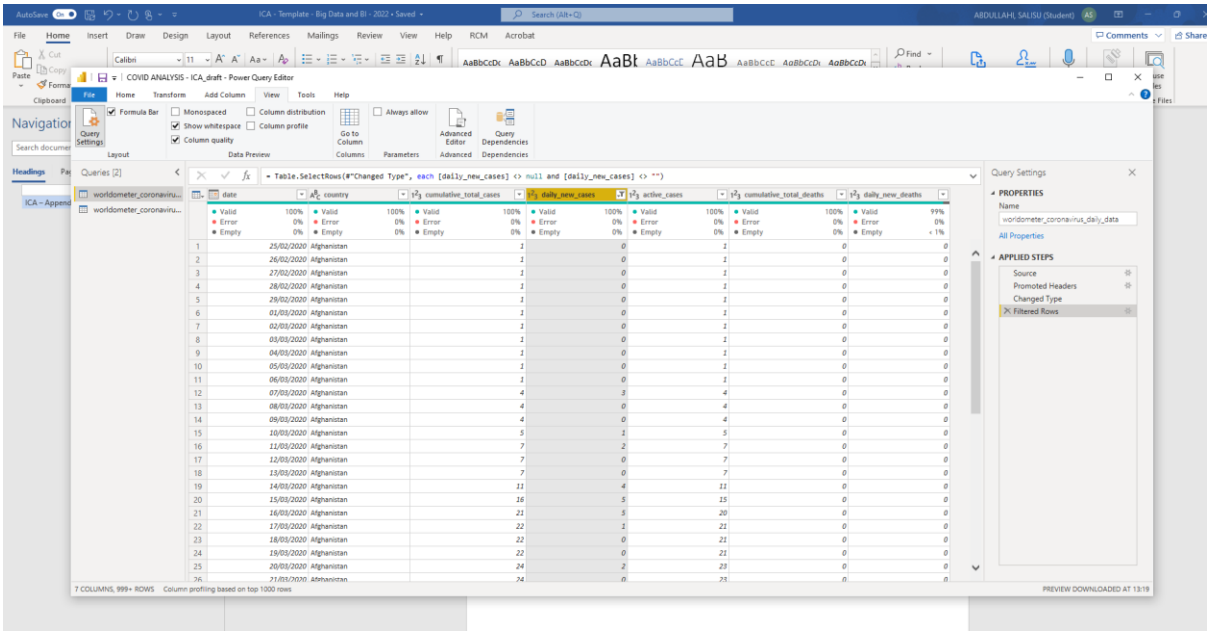


*Figure 2.4: Result of negative and null value removal on daily_new_cases column*

4. Step 3 was repeated, this time filter was applied on the **daily_new_death** column. Result shown below.

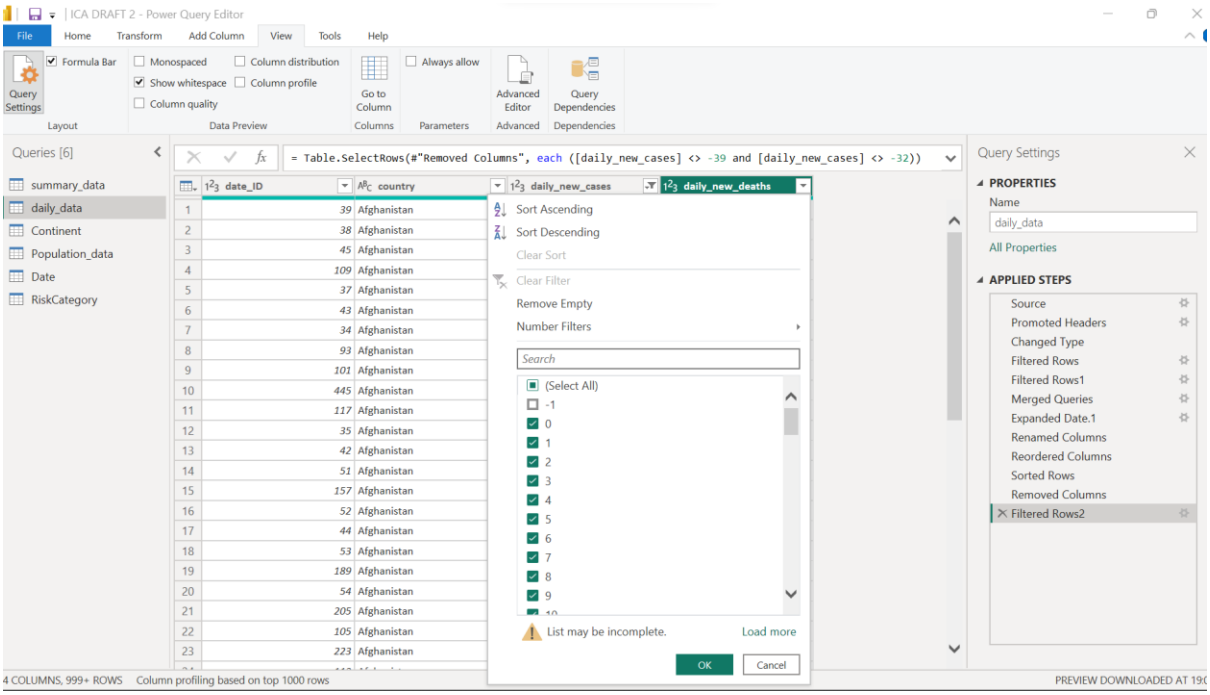



*Figure 2.5: Removal of negative values*

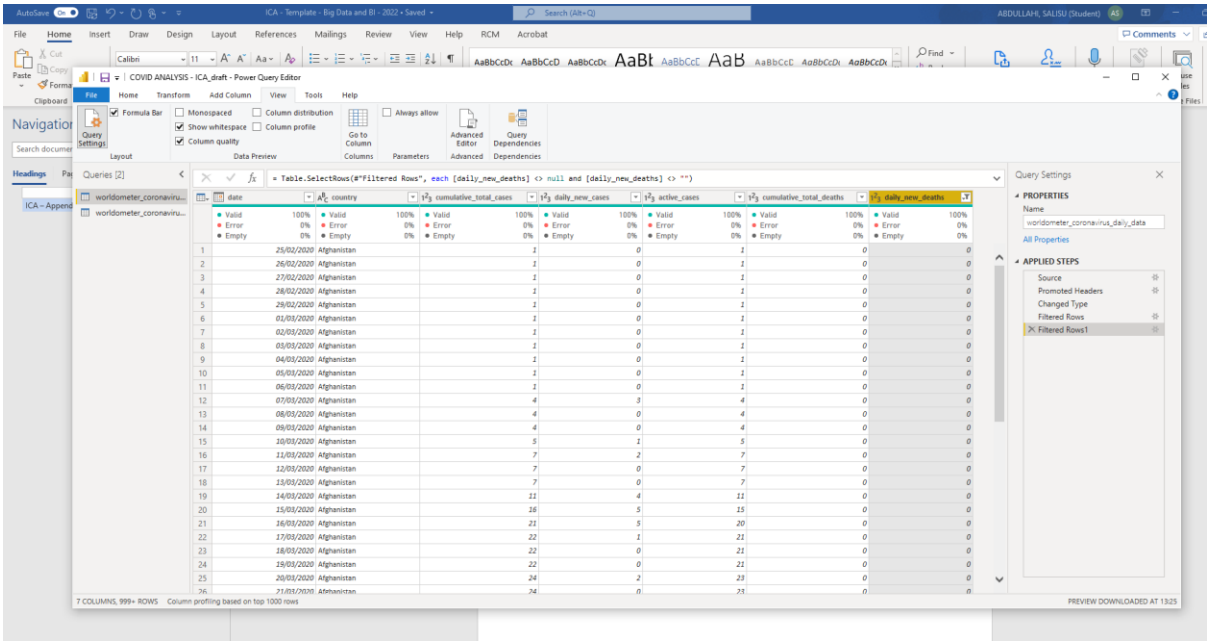


*Figure 2.6: Result of negative and null value removal on daily_new_deaths column*

5. I am now done cleaning the first data table (worldometer_coronavirus_daily_data) which I then renamed to **daily_data** table. Steps 2 to 4 was also applied to the second table (worldometer_coronavirus_summary_data) table which I renamed to **summary_data**. Worthy to note that the second table contained a column (serious_or_critical) with 10% more than 10 percent (36) of empty values in that column blank. Thus, before performing steps 2 to 4, I dealt with the missing values by identification, replacing null values in the column with 0. To do this, I selected the serious_or_critical column and then right clicked and selected the 'replace value' option to open the replace value dialogue box. On the box, I inputted 'null' in the 'value to find' space and '0' in the 'replace with' field. This process replaced all the null values with 0. Alternatively, the replace value option can be found on the **Home ribbon.** This was also applied to other columns with 10 percent empty column.

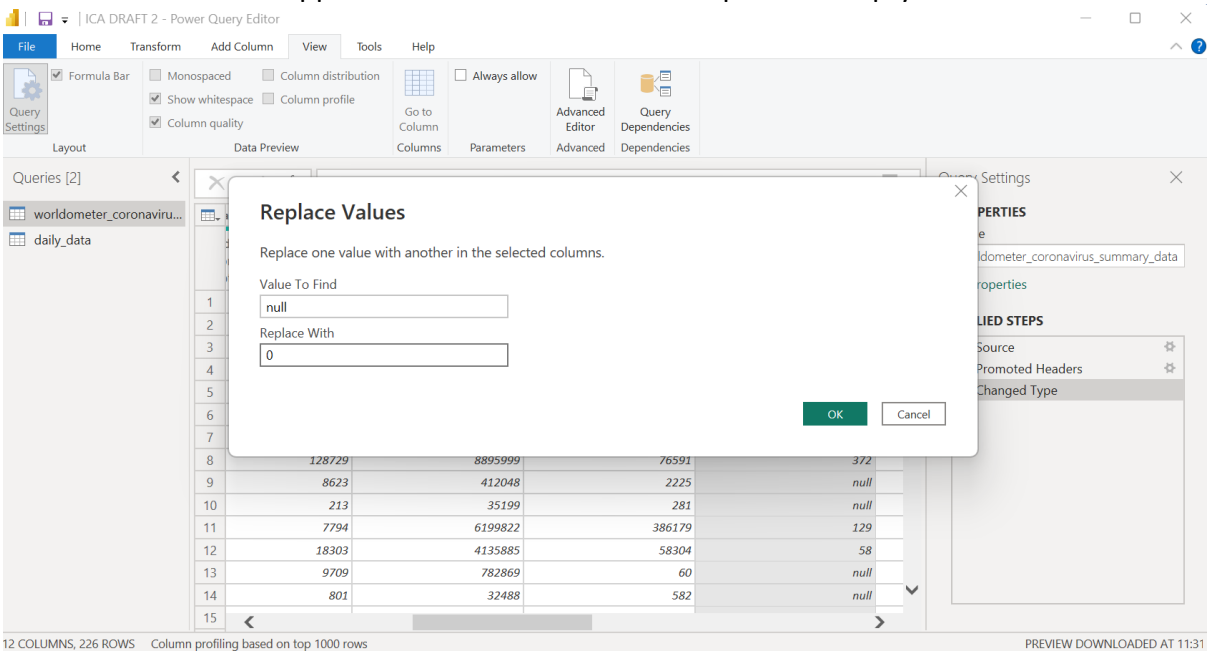



*Figure 2.7.1: Dealing with missing values by identification*

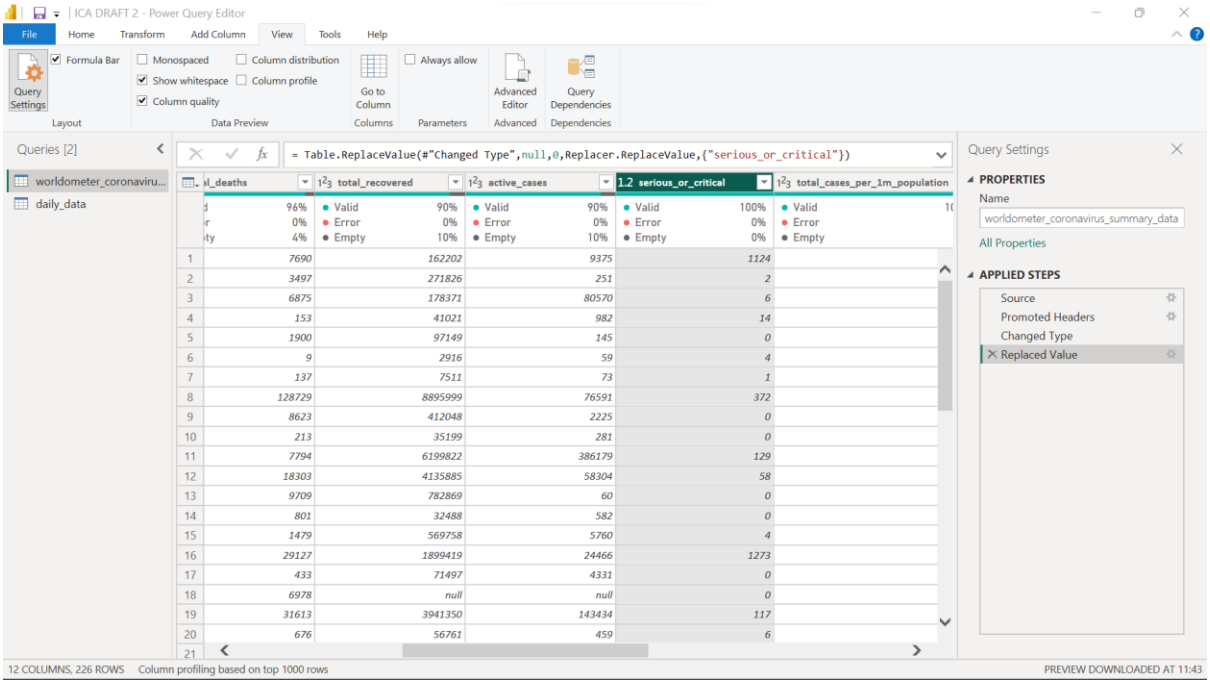



***Figure 2.7.2: Result of identification method***

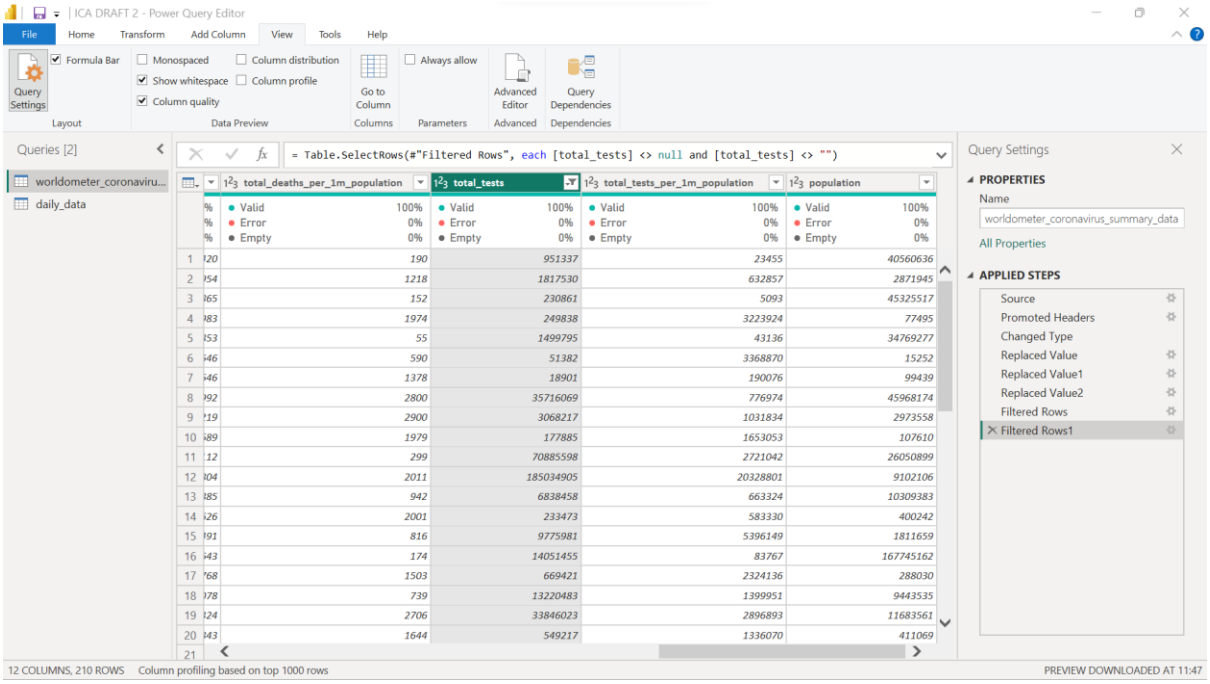



***Figure 2.8.1: Renaming the datasets***

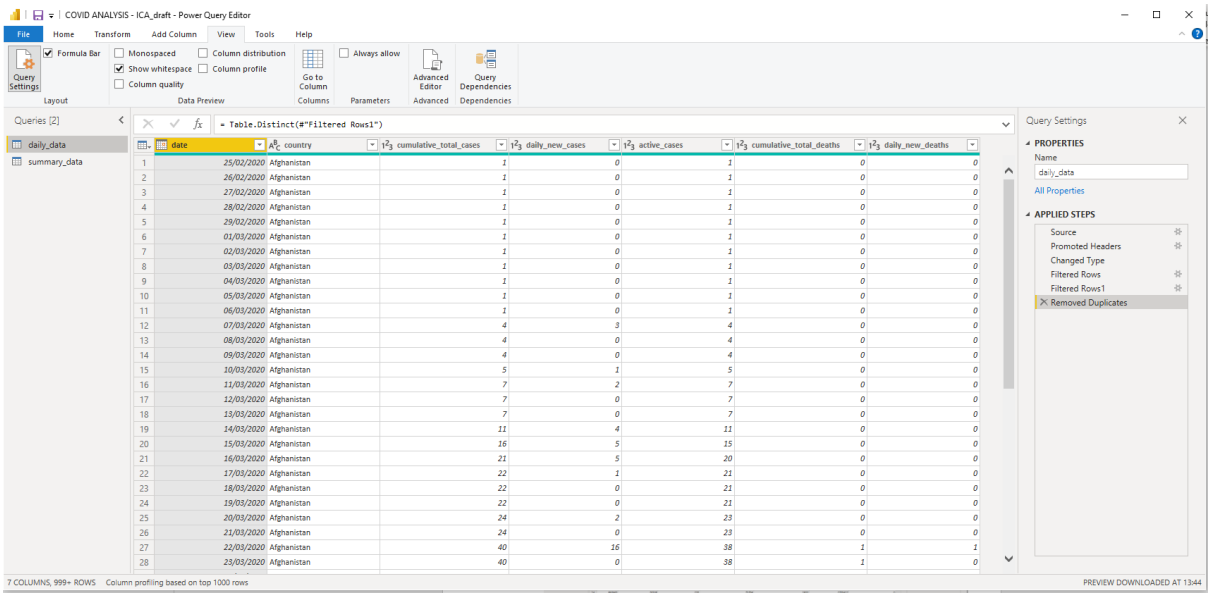



*Figure 2.8.2: screenshot of 'transform data' query showing renamed Datasets*

6. Now that I am done with cleansing for both data tables, we select the **close and apply** icon to apply all the changes made to the dataset table.

## BI Data Modelling via Star Schema - Facts and Dimensions.

1. Before dataset modelling begun, the data model state look was shown below. Existing relationship between the two tables was deleted as well. To achieve this, the relationship line was right-clicked and the 'delete' option was selected. This action can also be done by selecting the manage relationship icon from the home ribbon, ticking the relationship to be deleted from the dialogue box that opens and selecting delete.

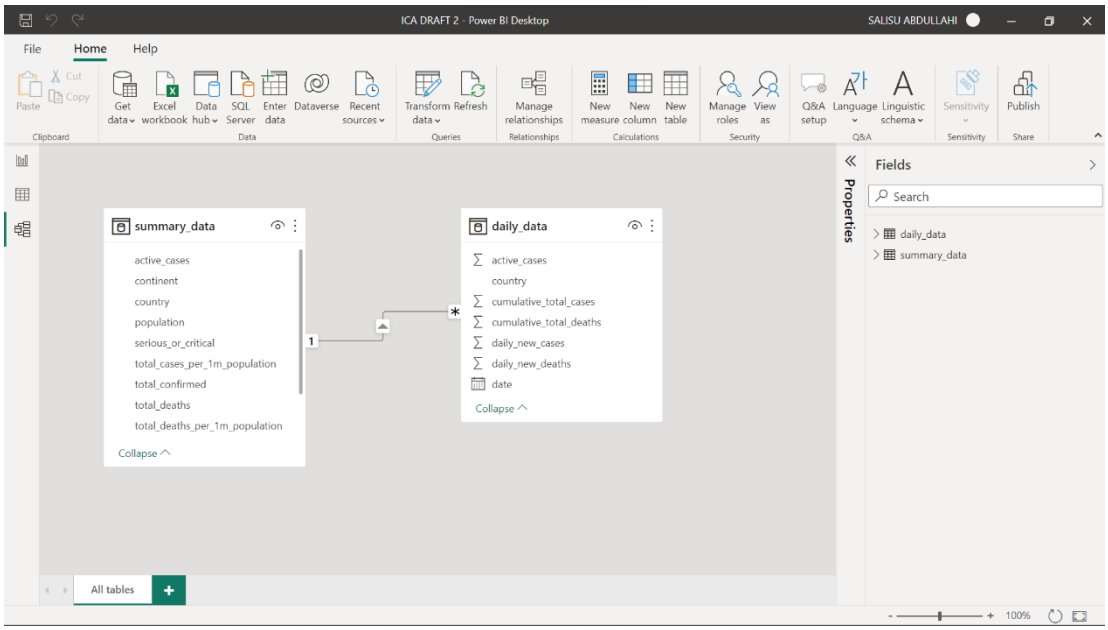
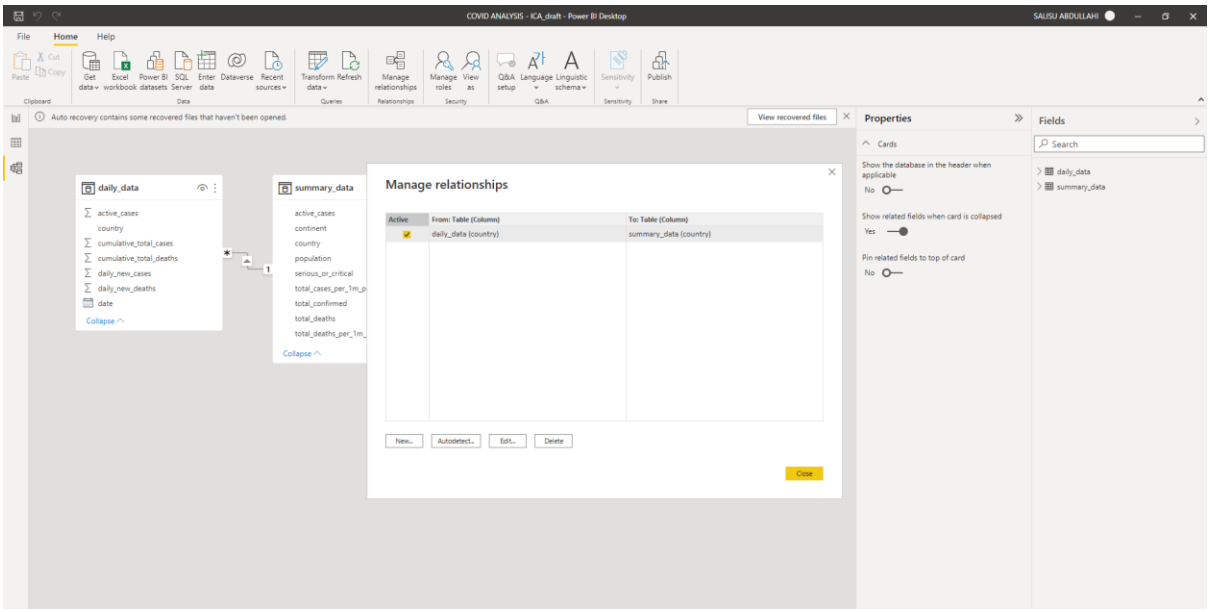
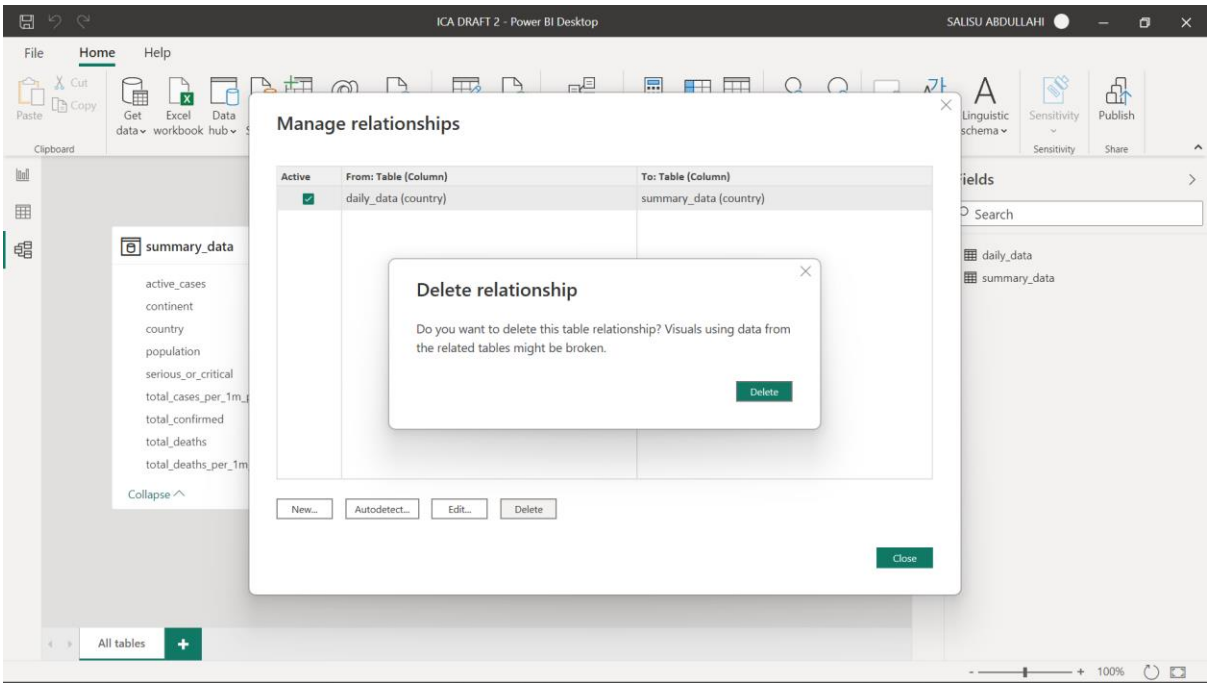



*Figure 3.0: screenshot of existing data model and relationship*


COVID ANALYSIS - ICA_draft - Power BI Desktop
SALISU ABDULLAHI
File
Home
Help
Manage relationships
daily_data
summary_data
Properties
Fields


ICA DRAFT 2 - Power BI Desktop
SALISU ABDULLAHI
File
Home
Help
Manage relationships
Active
From: Table (Column)
To: Table (Column)
daily_data (country)
summary_data (country)
Delete relationship
Do you want to delete this table relationship? Visuals using data from the related tables might be broken.
Delete
New...
Autodetect...
Edit...
Delete
Close
summary_data
active_cases
continent
country
population
serious_or_critical
total_confirmed
total_deaths
Collapse
Fields
Search
daily_data
summary_data
All tables
100%

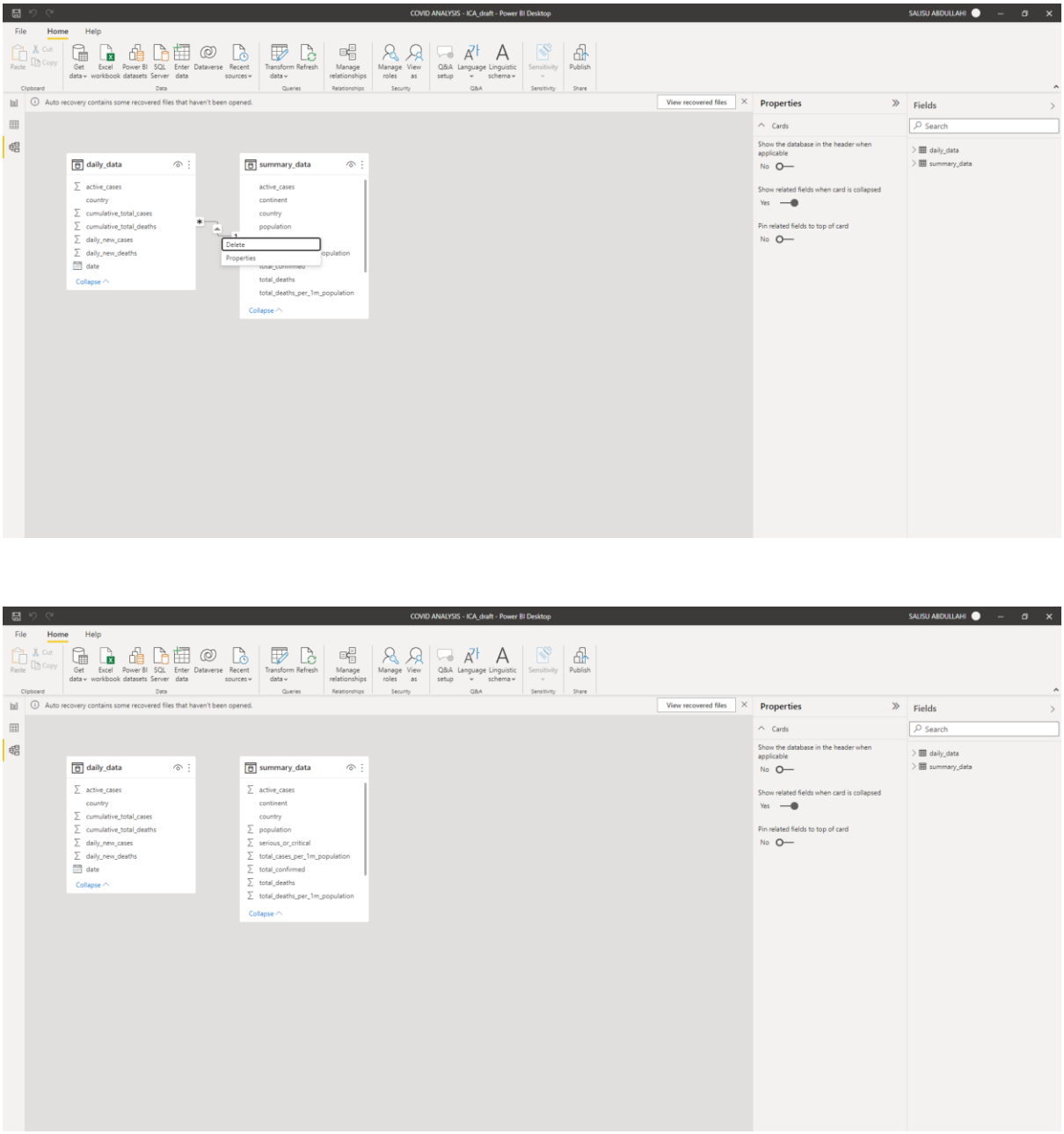



***Figure 3.1: screenshots showing data model relationship removal process***

2. When creating my data model for this analysis, I had to create dimension tables and fact tables. The aim of this is to de-congest and simplify my data table and make processing of my analysis easier. Dimension tables to be created include data that are contextual to the fact table.

## Dimension Tables

The first dimension table to be created was the **Continent** table. To do this, the first step was to click on transform data, then right click on the **summary_data** table and select duplicate. Then rename the duplicated table to Continent.

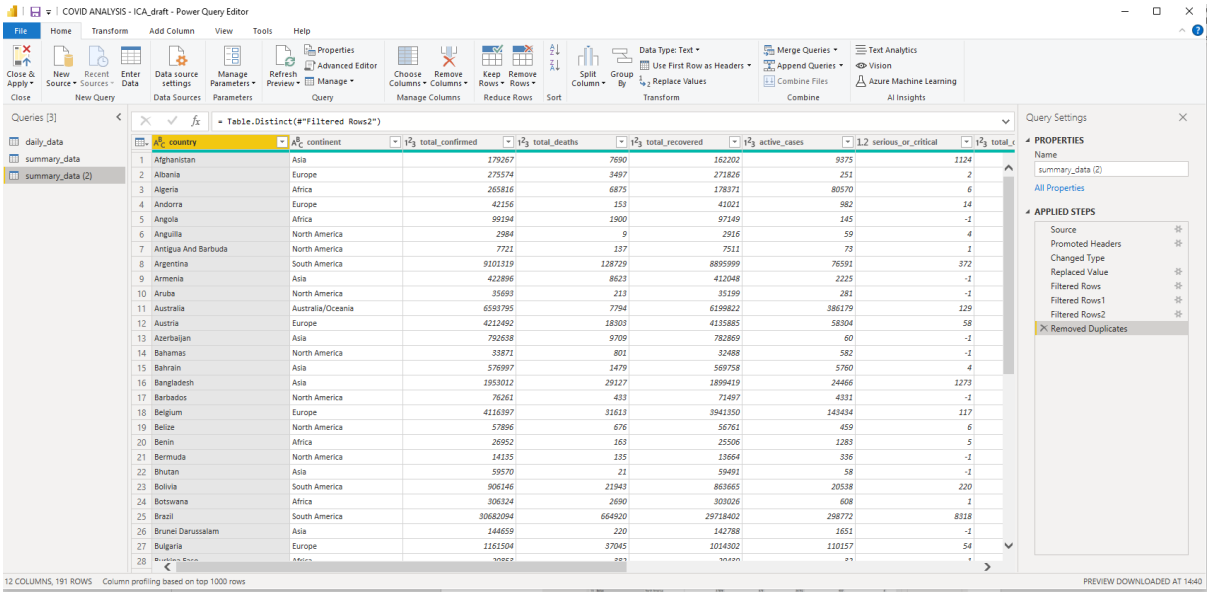



***Figure 3.2: screenshot showing duplicated data table***

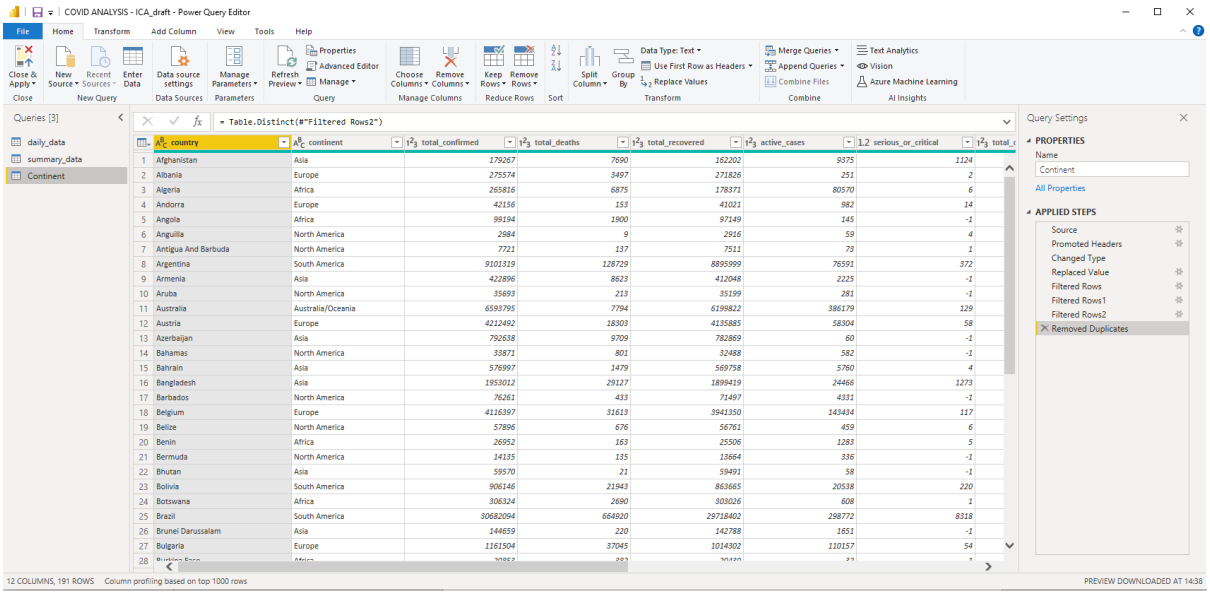



***Figure 3.3: screenshot showing renamed data table***

After renaming, the country and continent columns were selected. This is done by holding the control key and selecting both columns. Then other columns in the table were removed by right clicking on any of the selected columns and selecting the remove other columns option. I then remove duplicates from the data on the table by selecting the country column, right clicking on it and selecting remove duplicate.

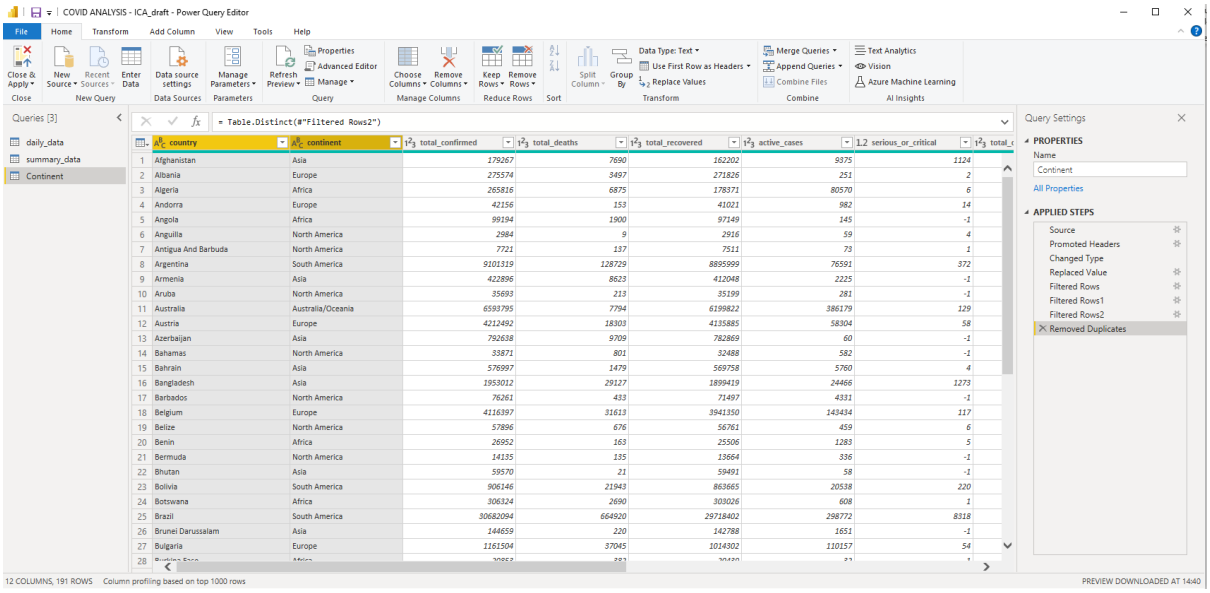



***Figure 3.4: screenshot showing data duplicate removal***

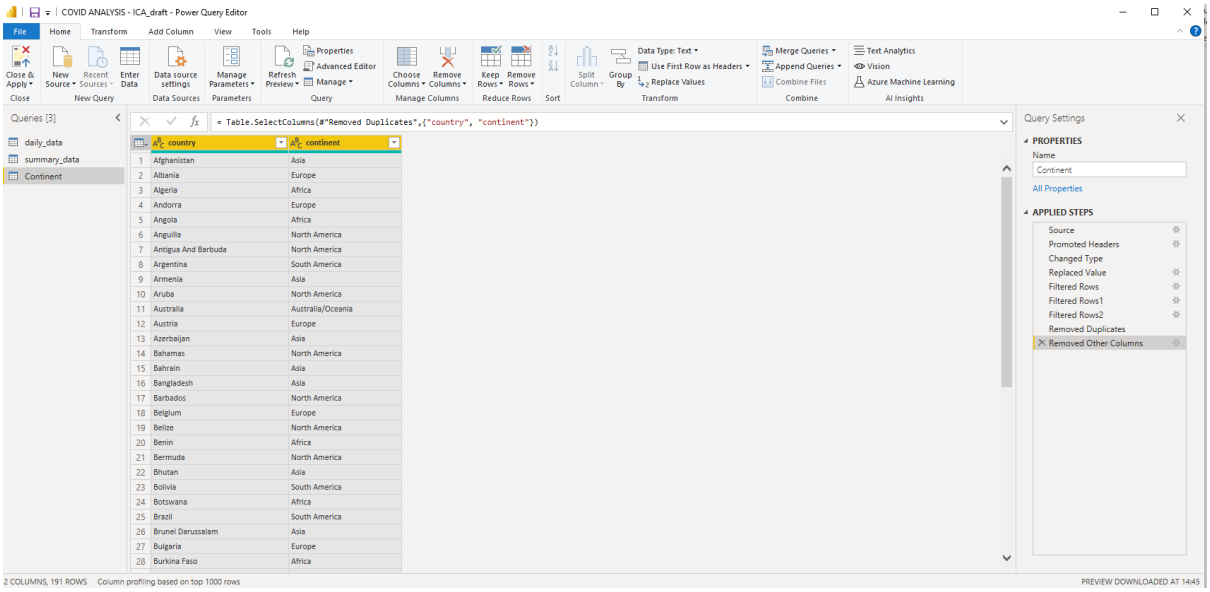



***Figure 3.5: screenshot showing other column removal***

We have successfully created the continent dimension table. We delete the continent column from the summary_data table. To do this, the continent column was selected, right clicked and he option ‘remove column’ was selected.

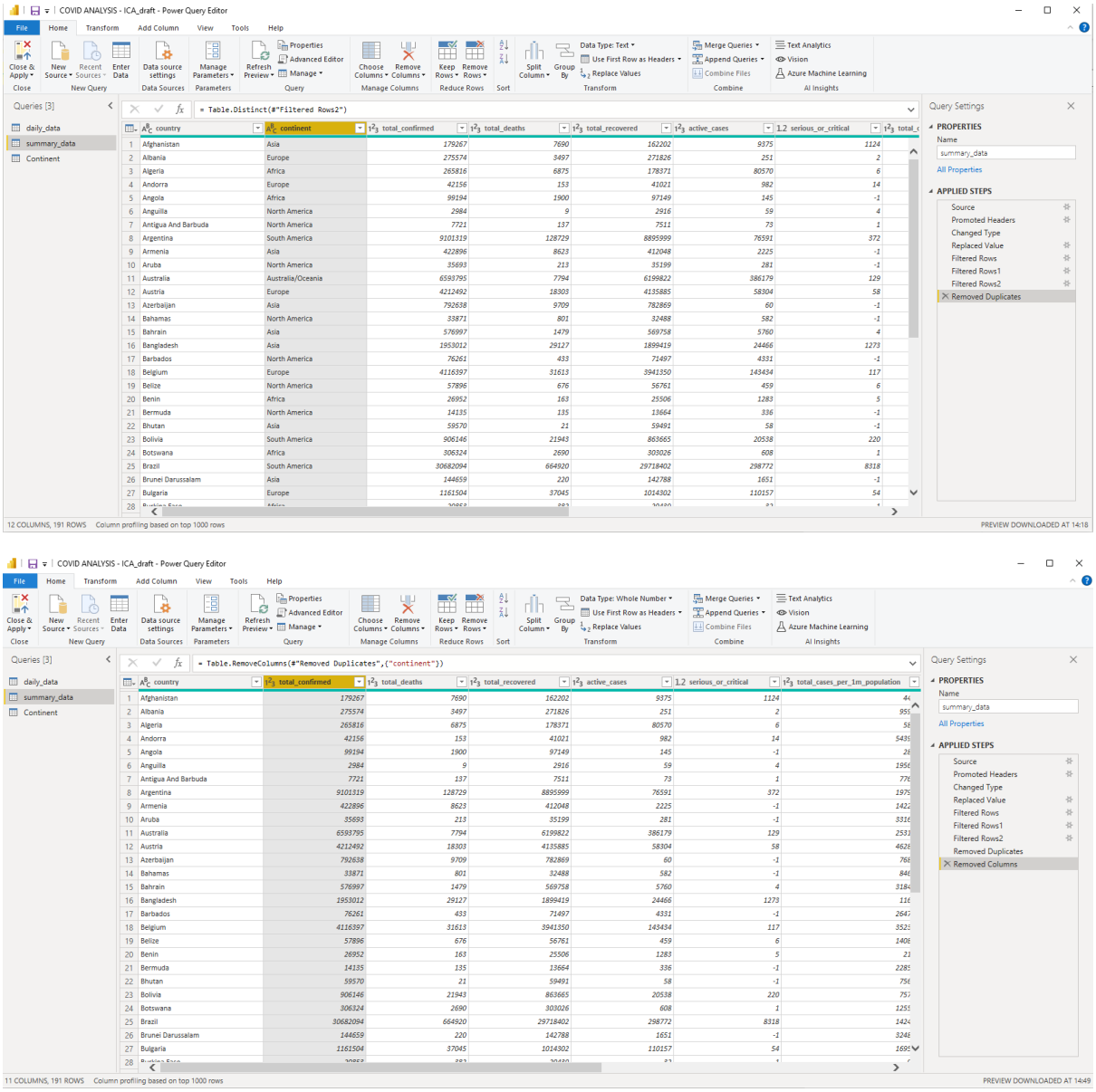
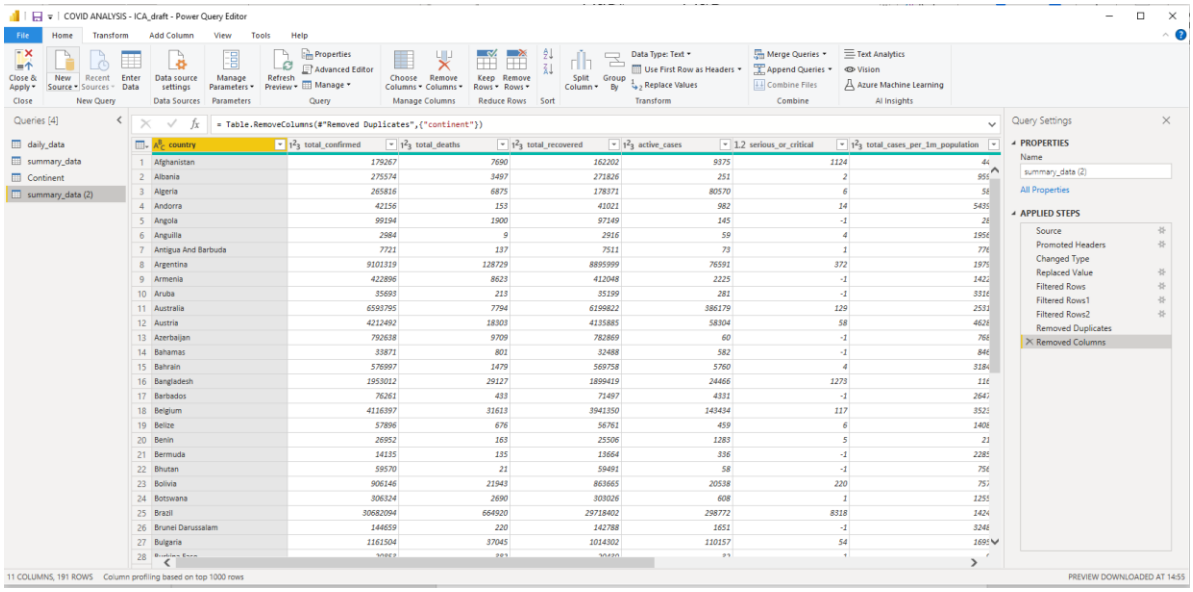
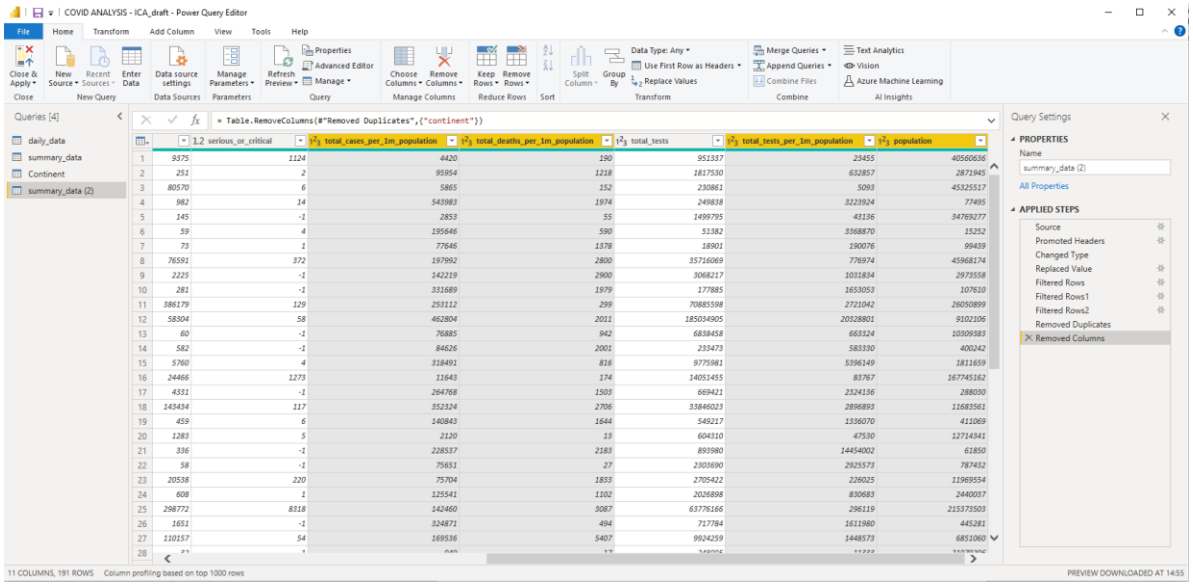



***Figure 3.6: screenshots showing continent column removal***

Next, I created a population_data dimension table. To do this, the previous dimension table steps were repeated but this time the columns selected were ‘Population’, ‘total_cases_per_1m_population’, ‘total_deaths_per_1m_population’, ‘total_tests_per_1m_population’ and ‘Country’.



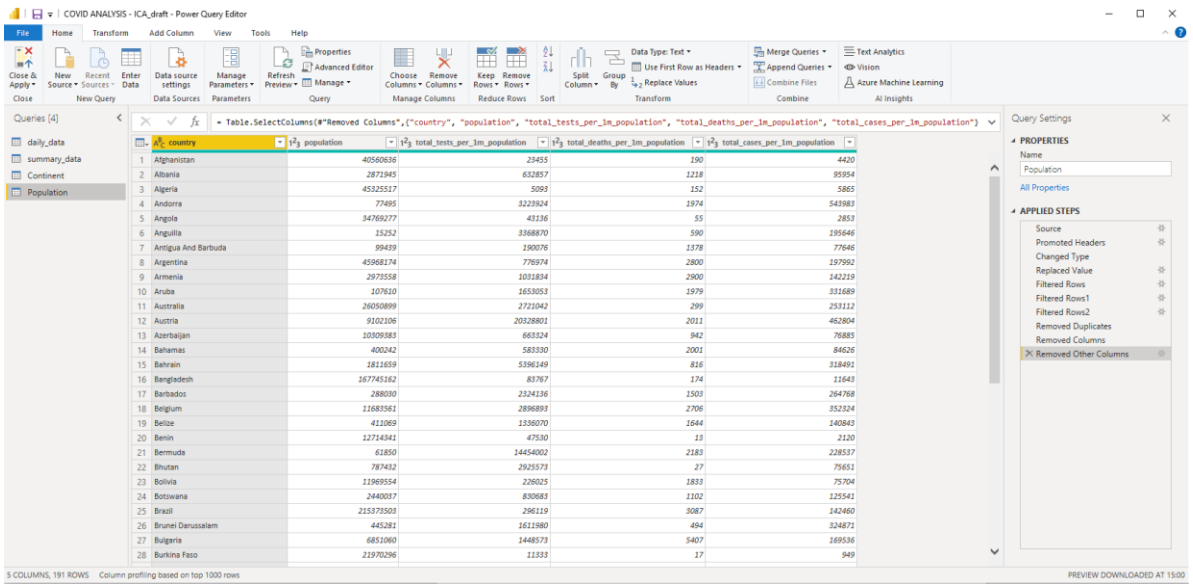




***Figure 3.7: screenshots showing Population table creation***

Upon creation of the population dimension table and deletion of ‘total_cases_per_1m_population’, ‘total_deaths_per_1m_population’ and ‘total_tests_per_1m_population’ columns from the summary_data table, the summary_data table looks like below.

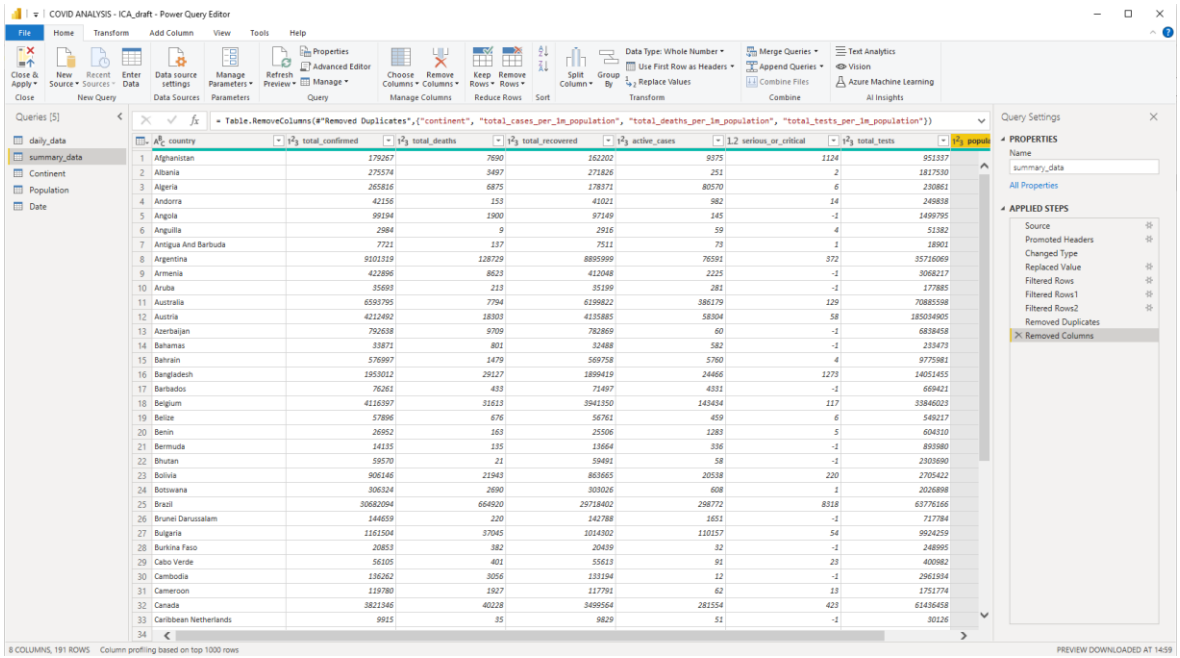



***Figure 3.8: screenshot of summary_data table after removal of population columns***

Next, I created a date dimension table. This was done by duplicating the daily_data table, selecting the date column and right click. Select ‘remove other column’ to remove the other columns.

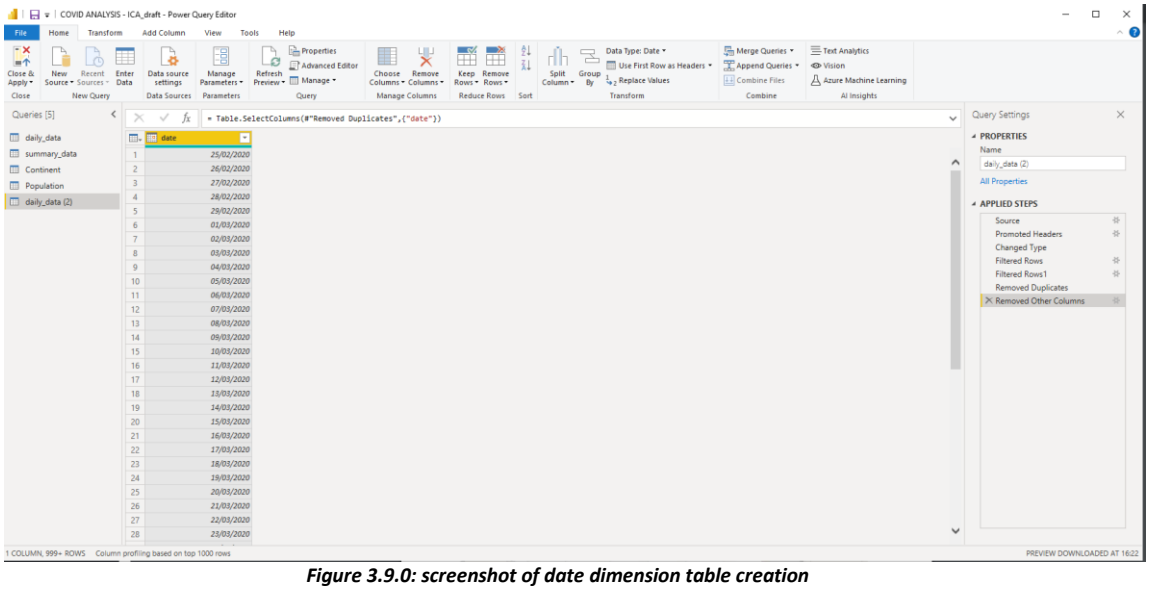



***Figure 3.9.0: screenshot of date dimension table creation***

Next step is to create an index for the date. This can be done by first removing duplicates of the date in your table. Right click on the date column and select remove duplicates. Next step is to sort the date in order of ascending. Click on the dropdown icon beside the date column and select ‘sort ascending’. I then clicked the index column dropdown and selected ‘from 1’ from the ‘add column ribbon’. Rename the table to ‘Date’.

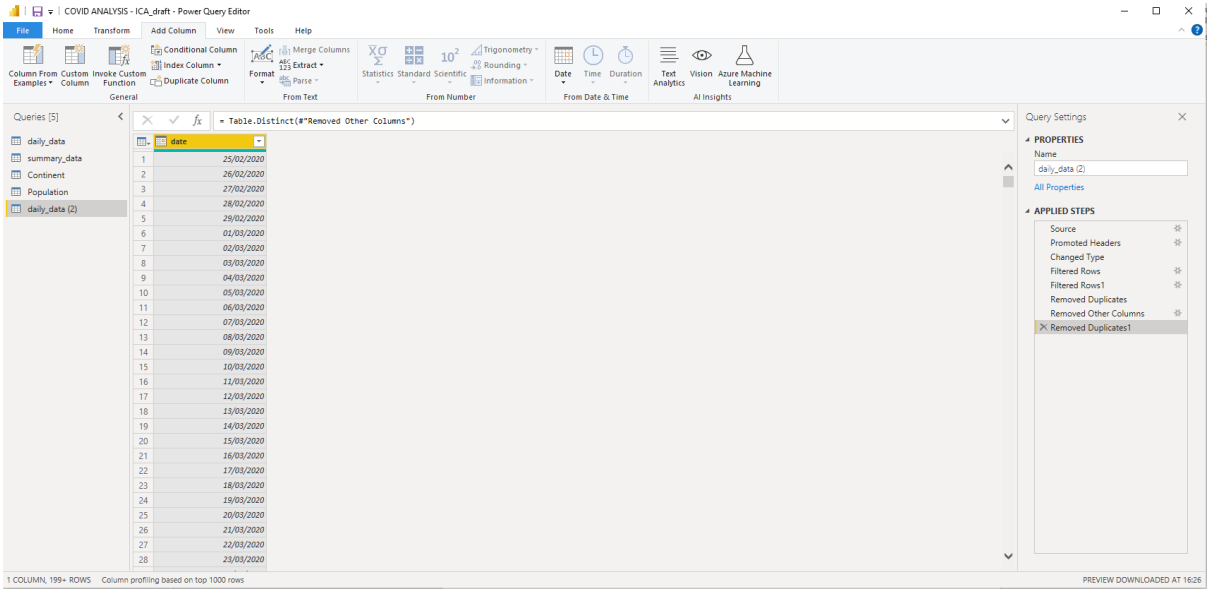



***Figure 3.9.1: screenshot showing date table after duplicate removal***

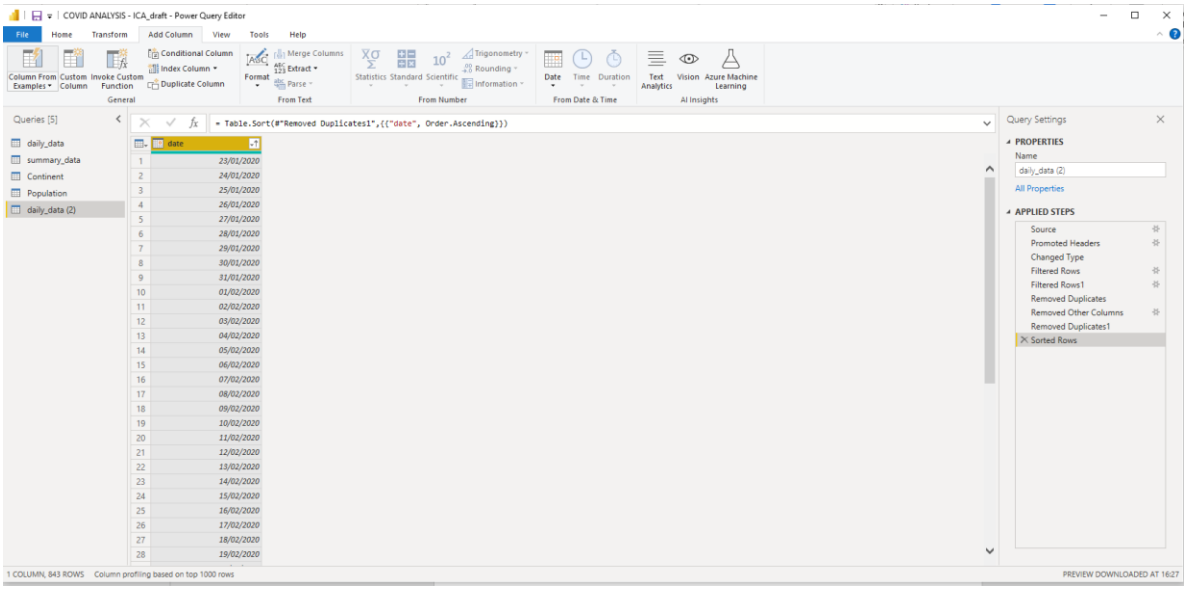



***Figure 3.9.2: screenshot showing date table after sorting rows***

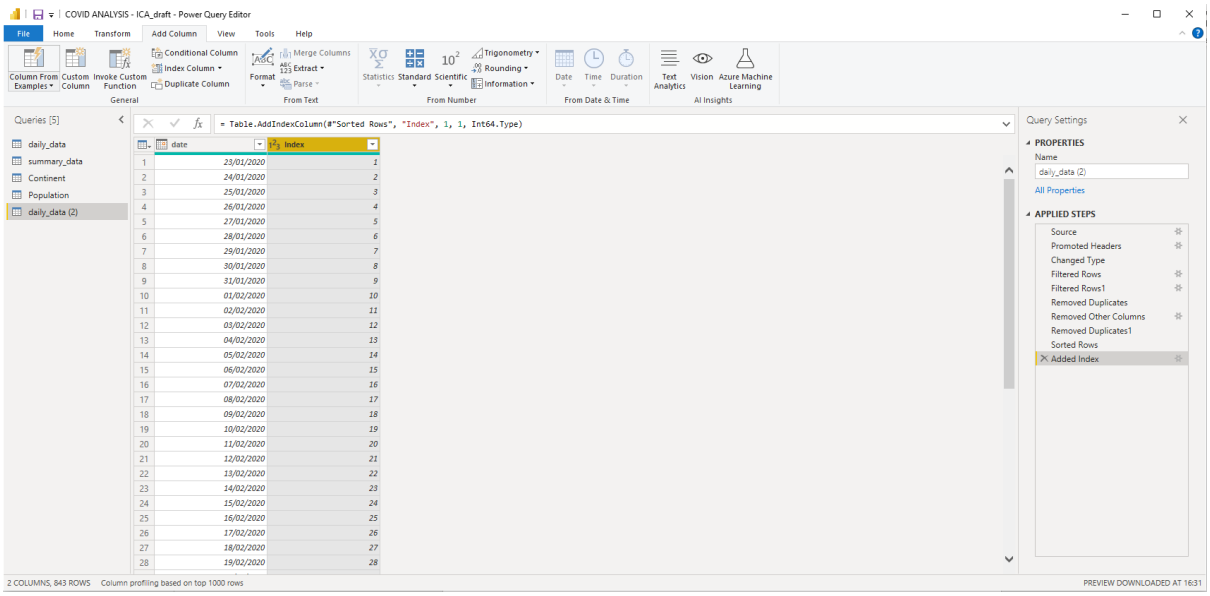


*Figure 3.9.1: screenshot showing date table after 'index column' addition*

Next, rename the index column on the date table to 'Date_ID' then merged the date table to the daily_data table. This was done by clicking on the merge queries icon from the 'home ribbon'. Selecting the date column from the daily_data and the date column from the date table and then click ok on the dialogue box. A new column was created. Click on the icon beside the new column and select '**date_ID'** only. Rename the new column to 'Date_ID'.

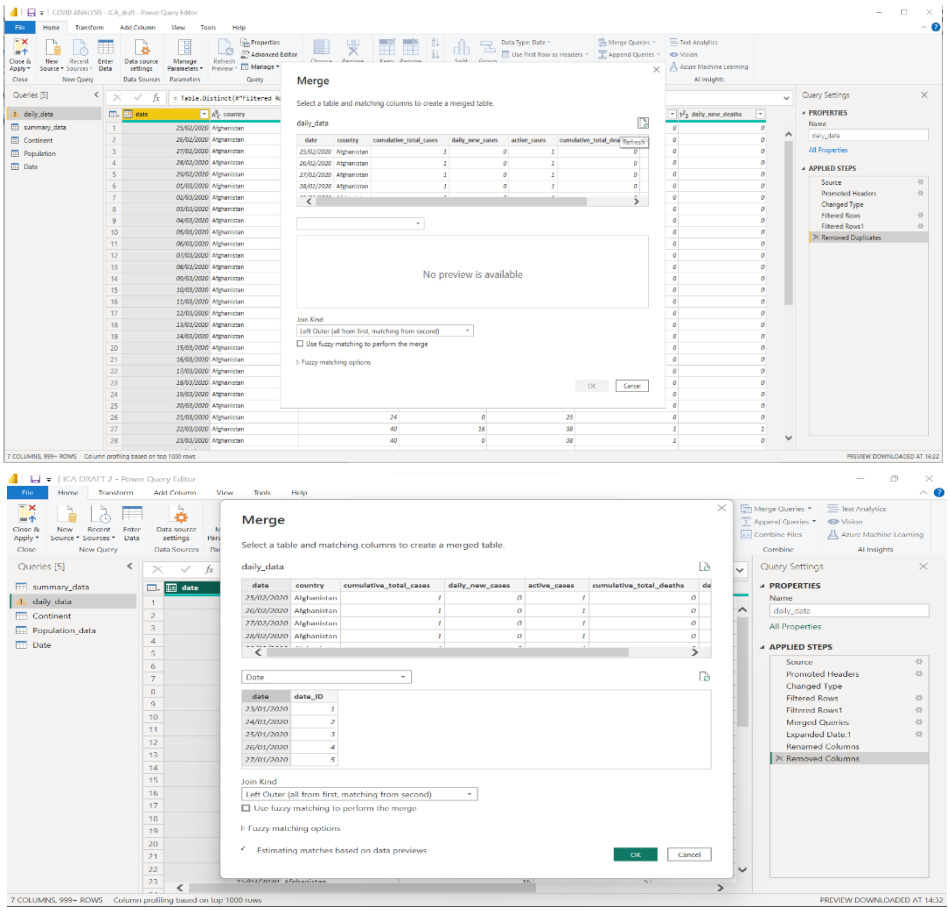


*Figure 3.9.2: screenshots showing table merge process*

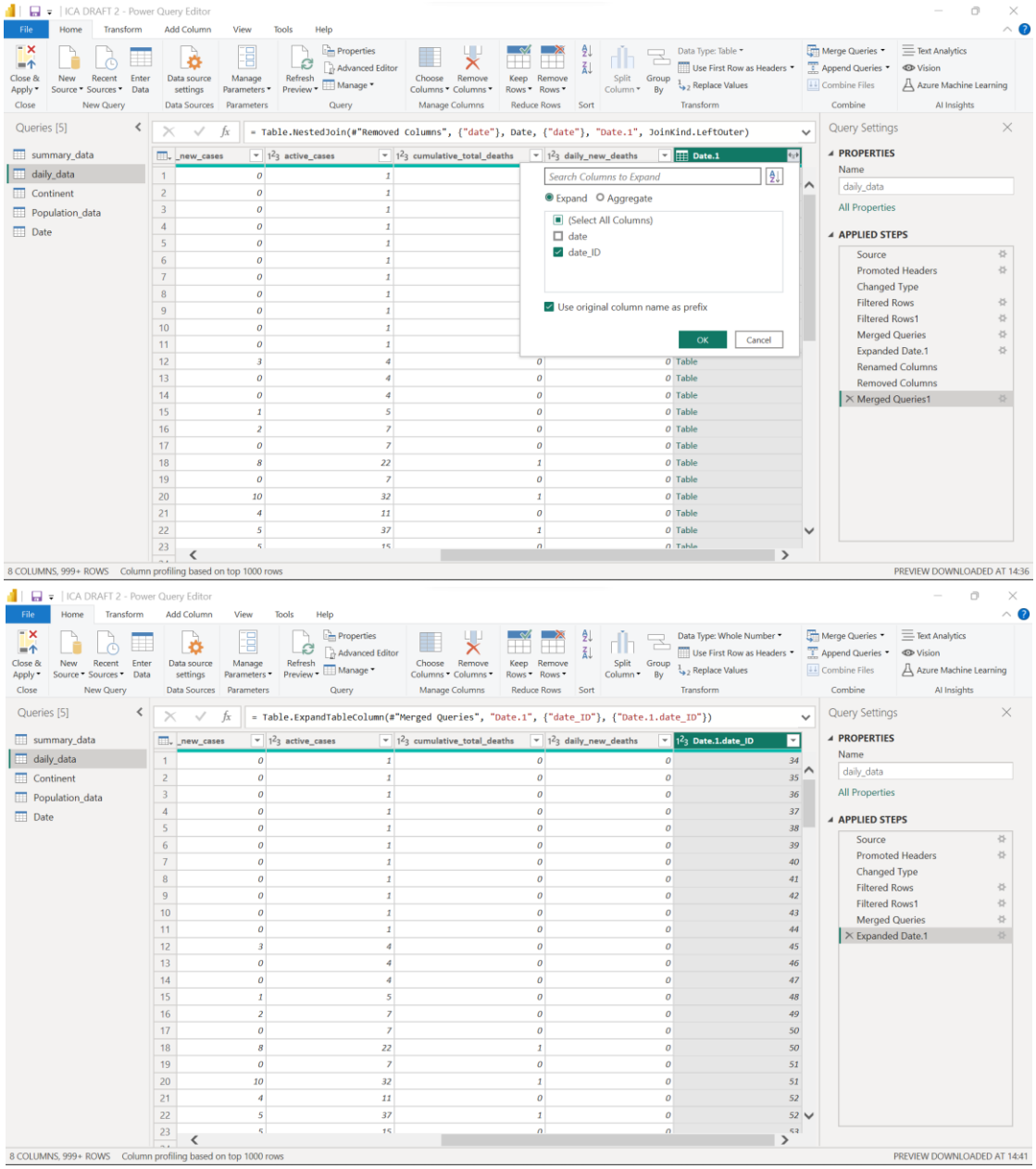



*Figure 3.9.3: screenshots of filtering table merged by date ID*

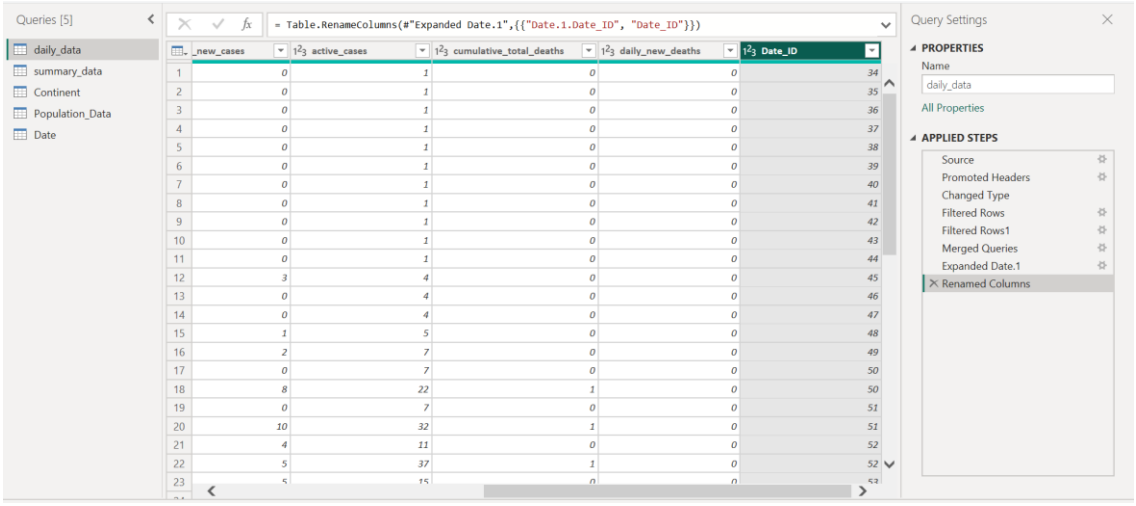



*Figure 3.9.4: screenshot of renamed date ID column*

The Date column on the daily_data column was then removed and the date_ID column was reordered to the first column. Sort the table by country.

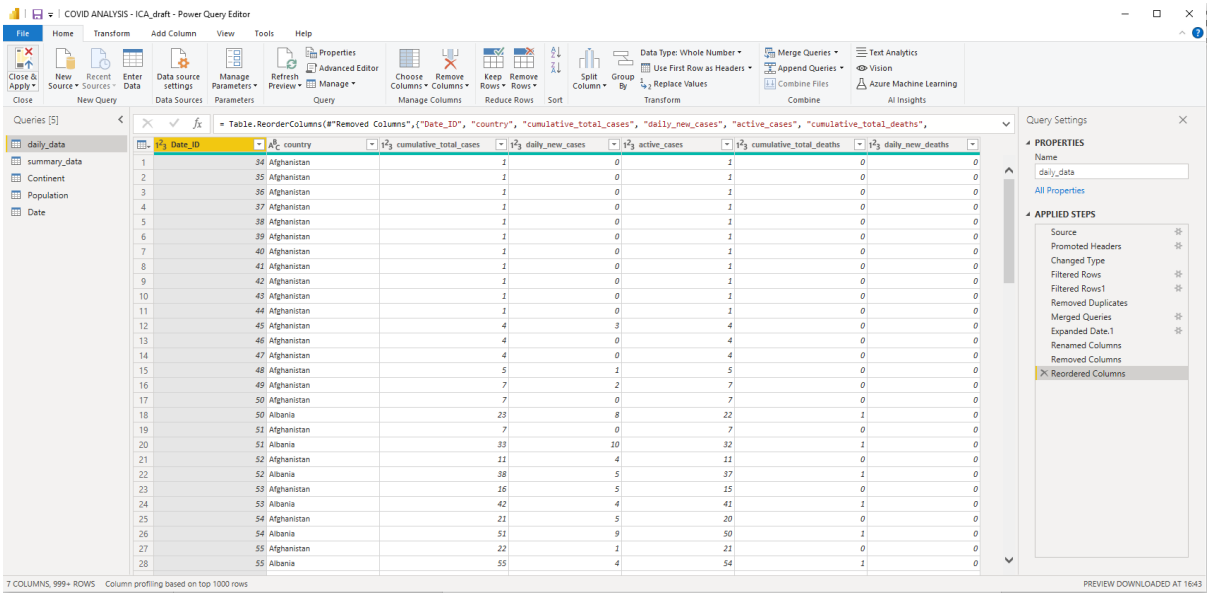



***Figure 3.9.5: screenshot showing daily_data table after date column removal and column reordering***

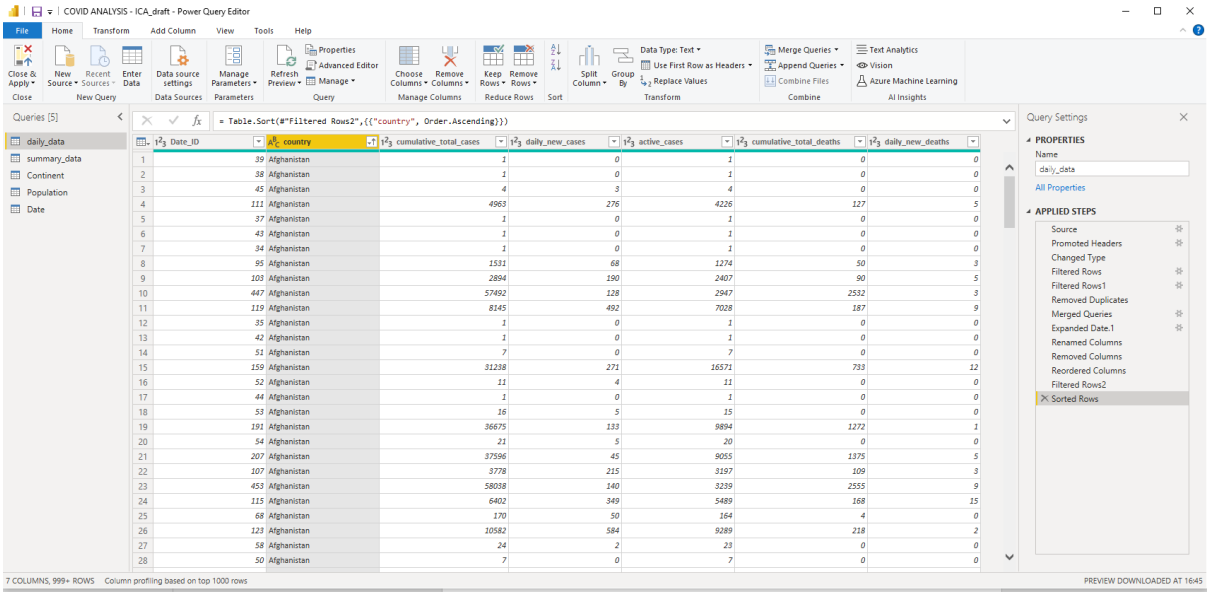



***Figure 3.9.6: screenshot showing daily_data table after sorting columns by Country***

We are now done creating dimension tables, we click on the 'close and apply' icon to apply changes.

## Fact Table

I used the daily_data table as my Fact Table. The reason behind my decision was because the daily_data table contains at least one column which are contextual or can be found on each of my dimension tables as well. My fact table included the following:

- **Date_ID:** representation of all the dates covering the period which this data was collected.
- **Country**: representation of all countries whose data was collected during this

exercise.

- **Daily_new_cases:** representation of confirmed cases of COVID-19 daily during the period of this study.
- **Daily_new_deaths:** representation of confirmed deaths due to COVID-19 daily during the period of this study.

All other columns in the daily_data table were removed as there were not relevant to the report. Click ‘Close & Apply’.

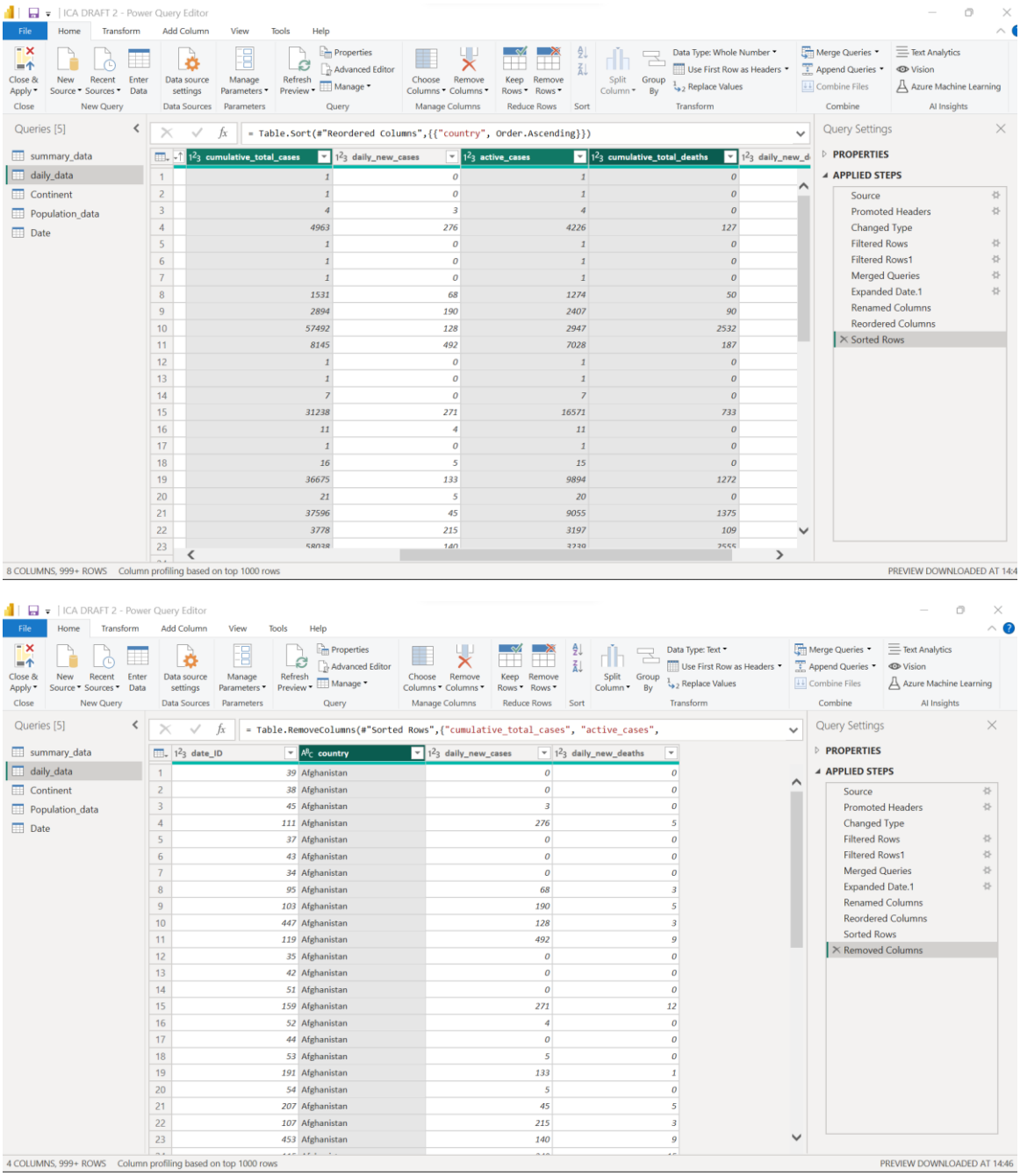



***Figure 4: screenshots showing removal of unwanted columns from the fact table***

The resulting data model was then created as a star schema as shown below. The date table is a dimension table with its keys(date_ID) connected to the daily_data table in a many to one relationship. The Continent table is dimension table connected to the daily_data table through the country column in a many-to-one relationship. The Population table is connected to the fact table through the country column in a many to one relationship. Lastly, the summary_data table is connected to the fact table through the country column in a many to one relationship.

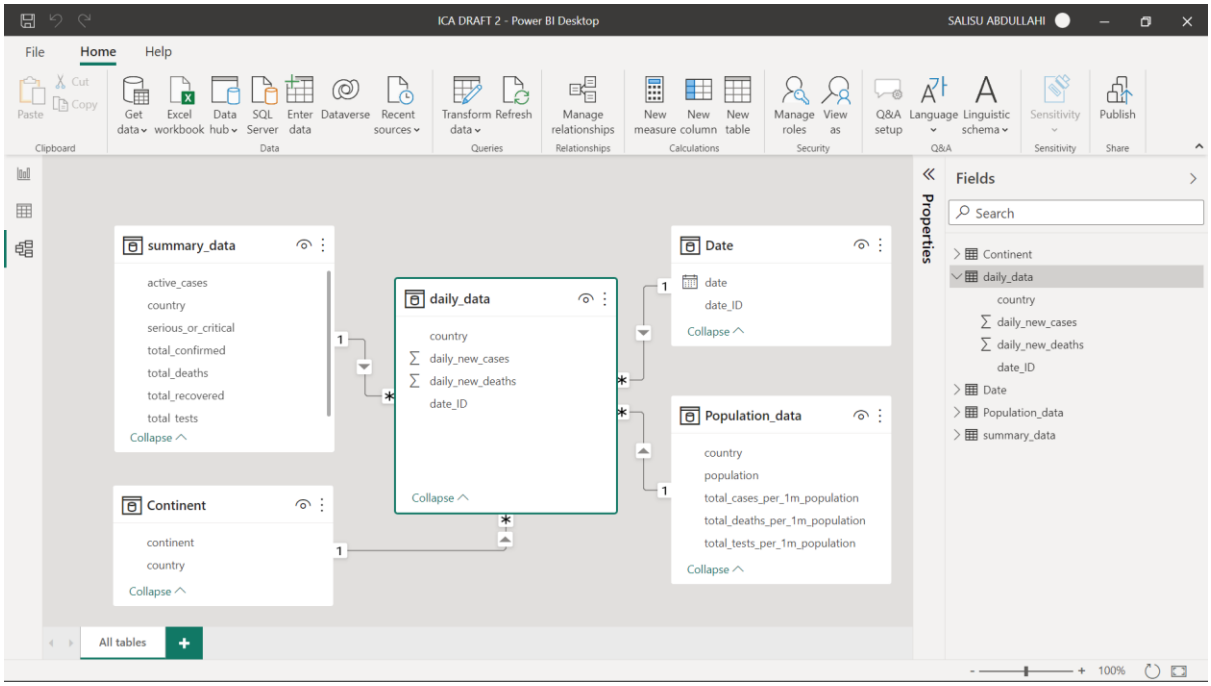



***Figure 5: screenshot showing Star Schema data model for Analysis***

## DAX and M Language

During this analysis, I needed to use DAX to create a calculated column which determines the mortality rate of COVID in each country. To do this, I selected 'new column' under calculations in the table tools ribbon after clicking on the Population_data Table under the Data tab. Next step was to name the column 'mortality_rate' and write the following formulas and press enter key to create a new column.

mortality_rate = Population_data[total_deaths_per_1m_population] / 1000000

where mortality_rate is the total deaths per 1 million of the population divided by 1 million. Now that a new column has been created. I formatted the mortality_rate data to be presented in percentages. This was done by clicking on the percentage symbol on under the 'column tools' ribbon.

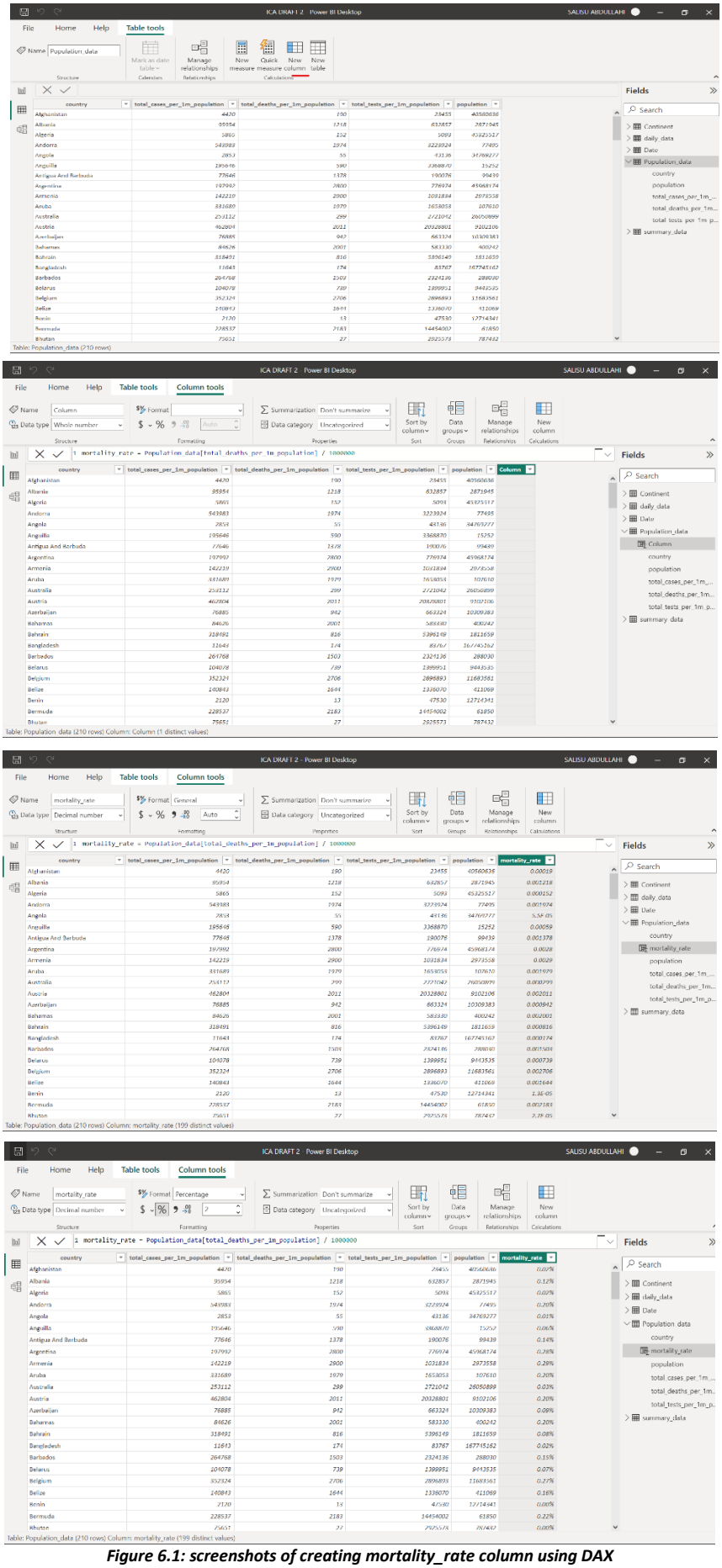


*Figure 6.1: screenshots of creating mortality_rate column using DAX*

Now that I have gotten the mortality, I decided to also get the infection rate in each country. Calculated columns were also created to determine the recovery rate in each country. To do this, same steps were applied to create a new column but below formula was used instead and the corresponding result is shown below.

Infection_rate = Population_data[total_cases_per_1m_population] / 1000000
Recovery_rate = summary_data[total_recovered] / summary_data[total_confirmed]
Where infection rate is total cases per 1 million of the population divided by 1 million and Recovery rate is the total recovered cases divided by the total confirmed cases (calculated under summary table)

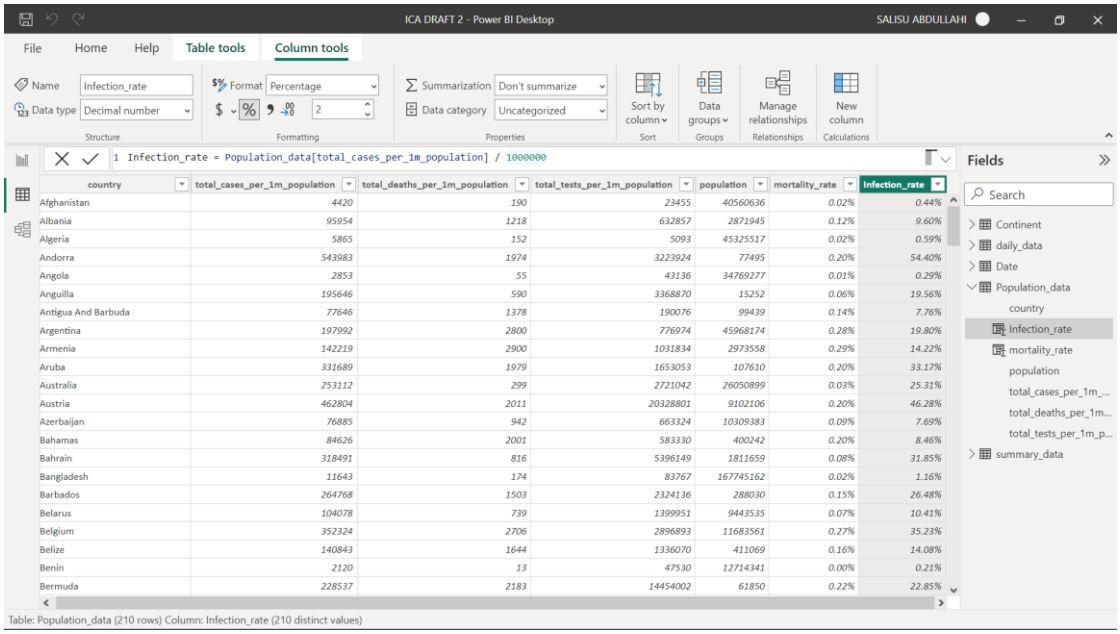



***Figure 6.2: screenshot of creating infection_rate column using DAX***

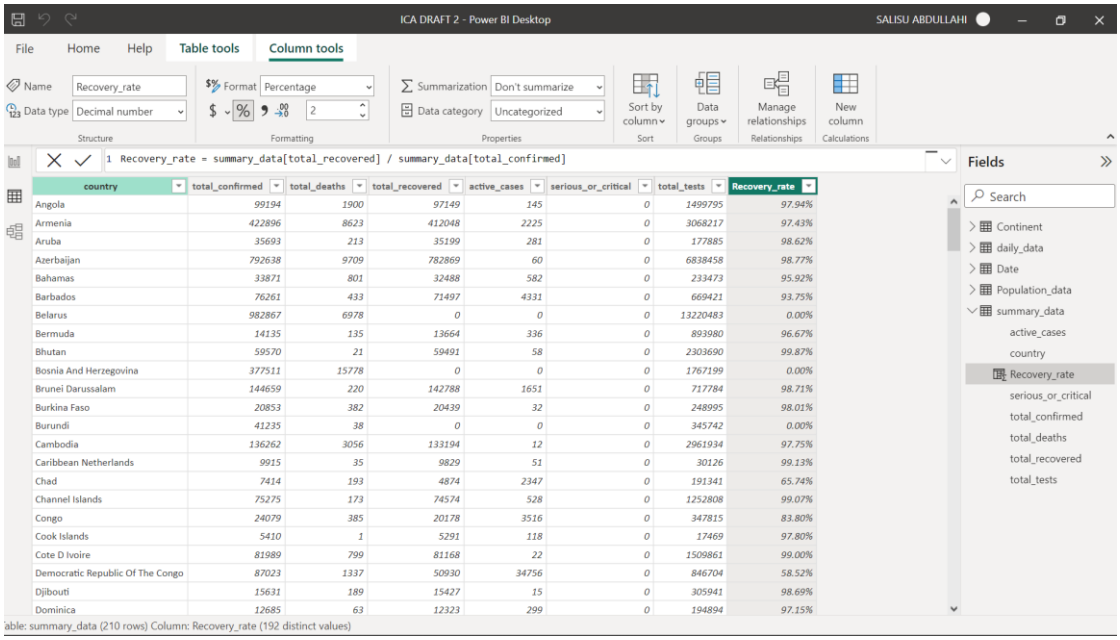



***Figure 6.3: screenshot of creating recovery_rate column using DAX***

Now that I have gotten the infection rate for each country, the next step was to find the average infection rate globally. To do this, I created a calculated measure which took the total of the infection rates of all the countries under analysis and find the average. I clicked on the population_data table and selected on the new measure icon from the 'table tools' ribbon and wrote the following in the formula box.

Global_Average_Infection = AVERAGE(Population_data[Infection_rate])

Where the global average infection rate is the average of the infection rate total.

The resultant value was also shown in percentages using same steps I used on the calculated column (clicking on the percentage symbol on under the 'column tools' ribbon).

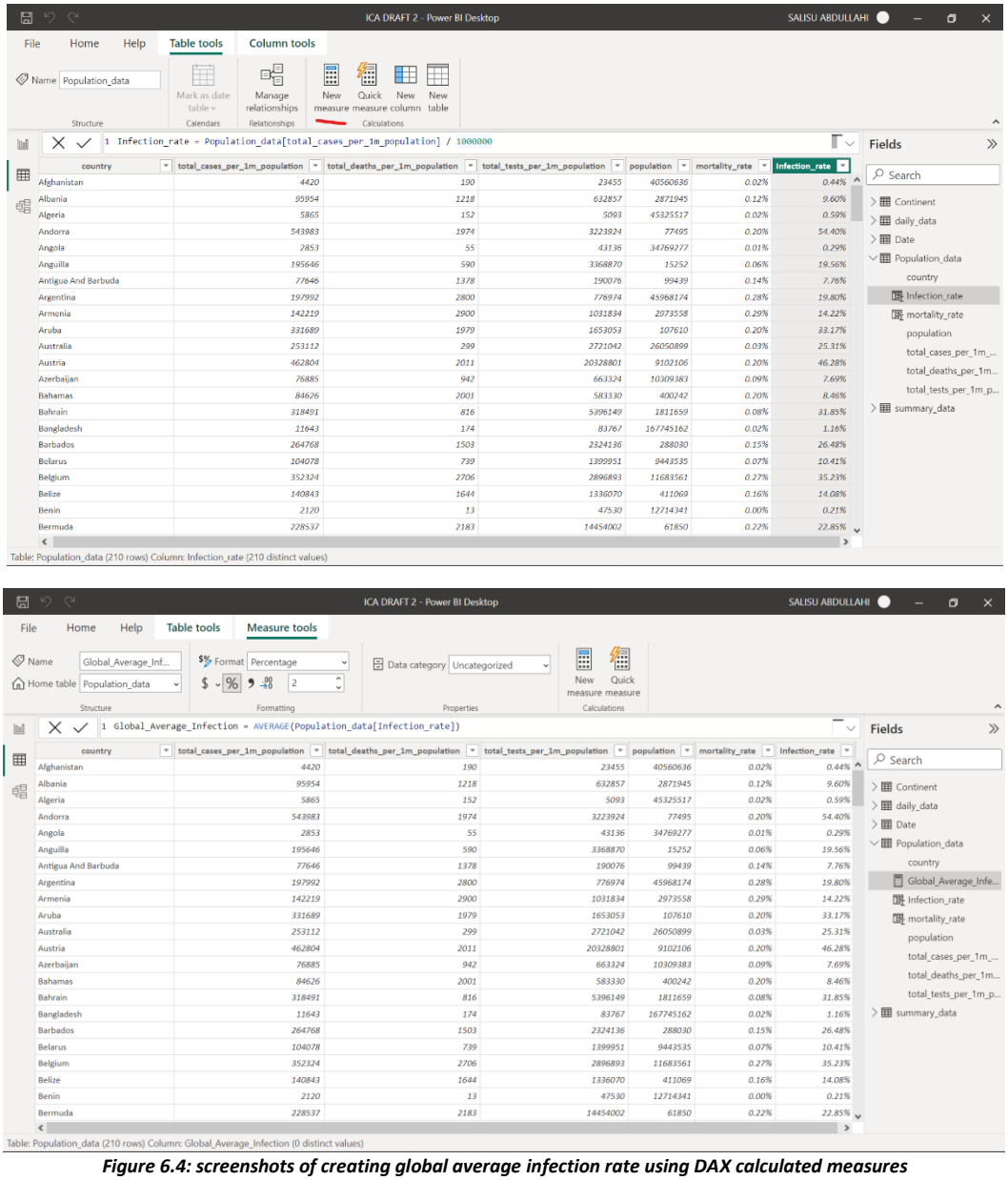



***Figure 6.4: screenshots of creating global average infection rate using DAX calculated measures***

Calculated measures were also created under the daily_data table to determine the number of covid confirmed infection cases, number of covid deaths globally, and global number of recovered cases. Same steps as above were applied, and the corresponding formulas used are shown below.

Total_Infected = SUM(daily_data[daily_new_cases])
total_deaths = SUM(daily_data[daily_new_deaths])
total_recovered = SUM(summary_data[total_recovered])

**where**

- Total infected is the sum of daily new cases of covid
- Total Deaths is the sum of daily new deaths due to covid
- Total recovered is the sum of the total recovered cases of covid

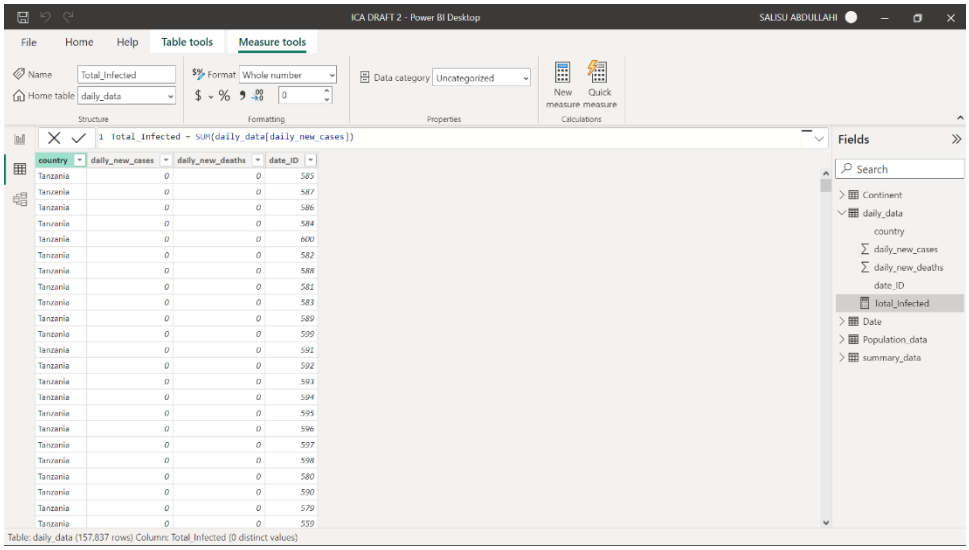



***Figure 6.5: screenshot showing total infected measure created using DAX calculated measures***

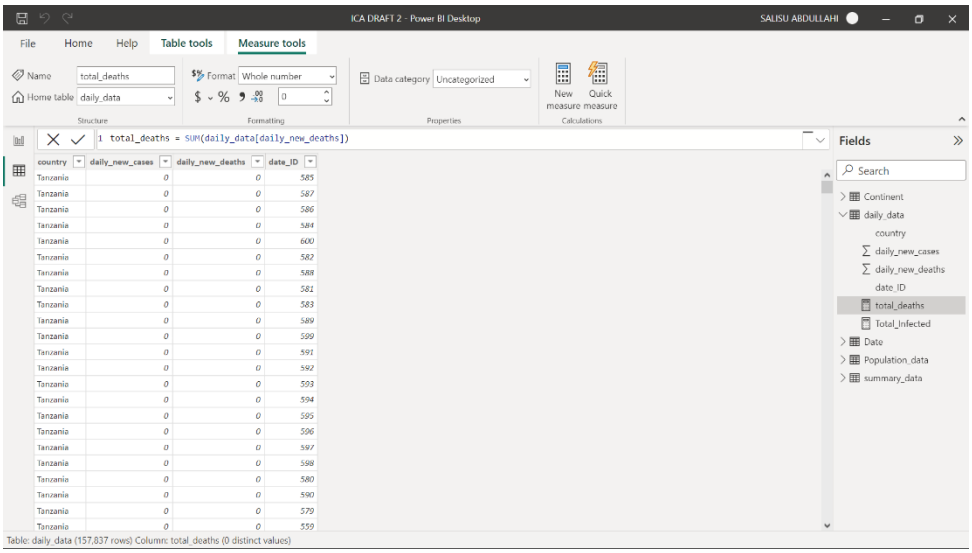



***Figure 6.5: screenshot showing total deaths measure created using DAX calculated measures***

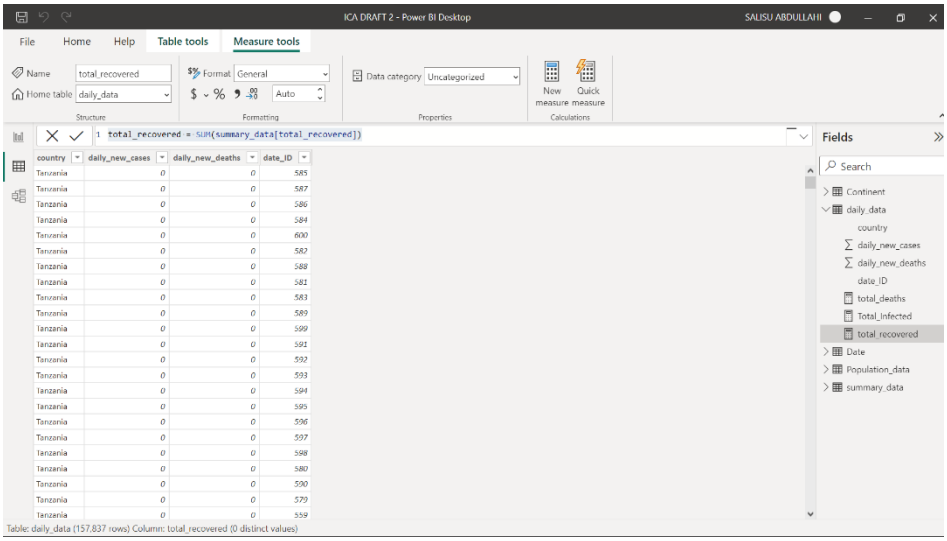



***Figure 6.5: screenshot showing total recovered measure created using DAX calculated measures***

## M. Language.

Now that I have created the respective calculated columns and measures, my next step is to outline countries based on their infection rates in my report to provide insight. To do this, I used M. Language to create a risk category based on infection rate.

This was done by first selecting 'Get data' and clicking on 'Blank Query' from the menu.

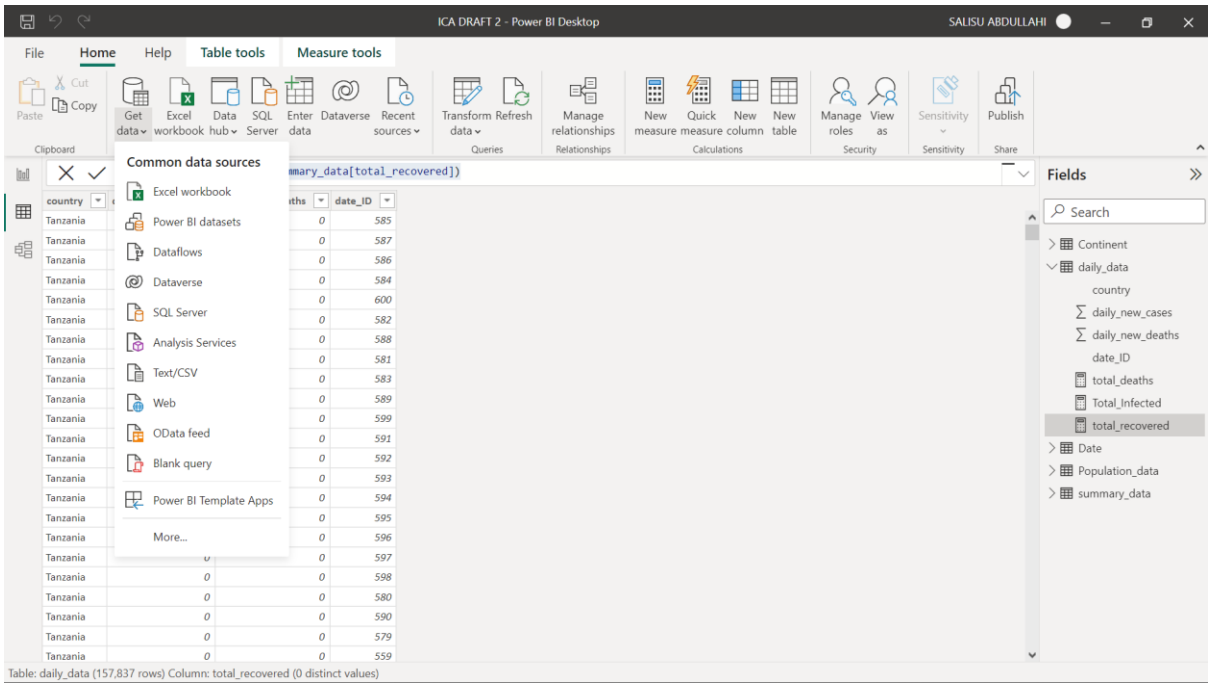



***Figure 7.0: screenshot showing 'Get Data' and 'Blank Query' selection***

The next step was to select 'Advanced Editor' from the dialogue box that opened.

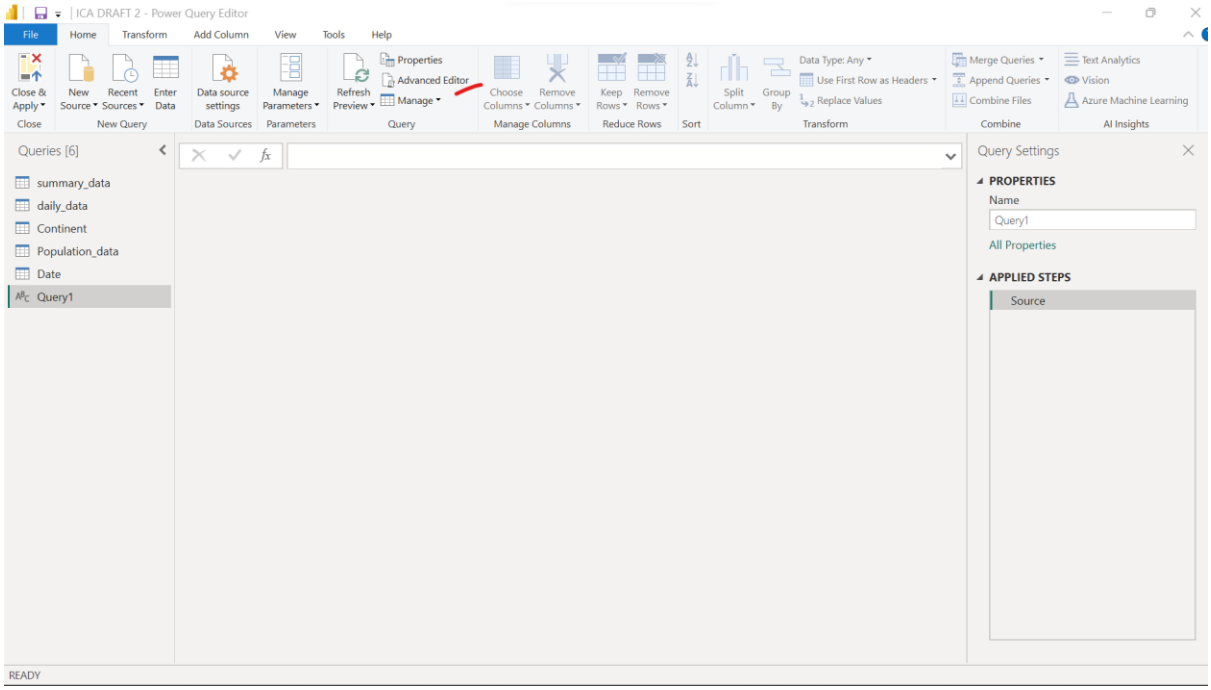



***Figure 7.1: screenshot showing 'advanced editor' selection***

The Table.fromrecord function was then called and the following below instructions were inputted and done was clicked. Rename the table to Risk Category and Click on close and apply.

```
Source = Table.FromRecords ({
    [Category = "0-15%", Lower Level = 0, Upper Level = 0.15],
    [Category = "15-30%", Lower Level = 0.15, Upper Level = 0.30],
    [Category = "30-50%", Lower Level = 0.30, Upper Level = 0.50],
    [Category = "50-70%", Lower Level = 0.50, Upper Level = 0.70],
    [Category = "70-100%", Lower Level = 0.70, Upper Level = 1]
})
```

**Where**

- Table.FromRecords instructs the programme to create a table.
- Category, Lower Level and Upper Level are columns created in the table.

Next step, I changed the data type on lower and upper level to decimals. This was done by selecting transform data and then clicking on the icon beside the lower level and upper level icon and selecting decimal number. Click close and apply.

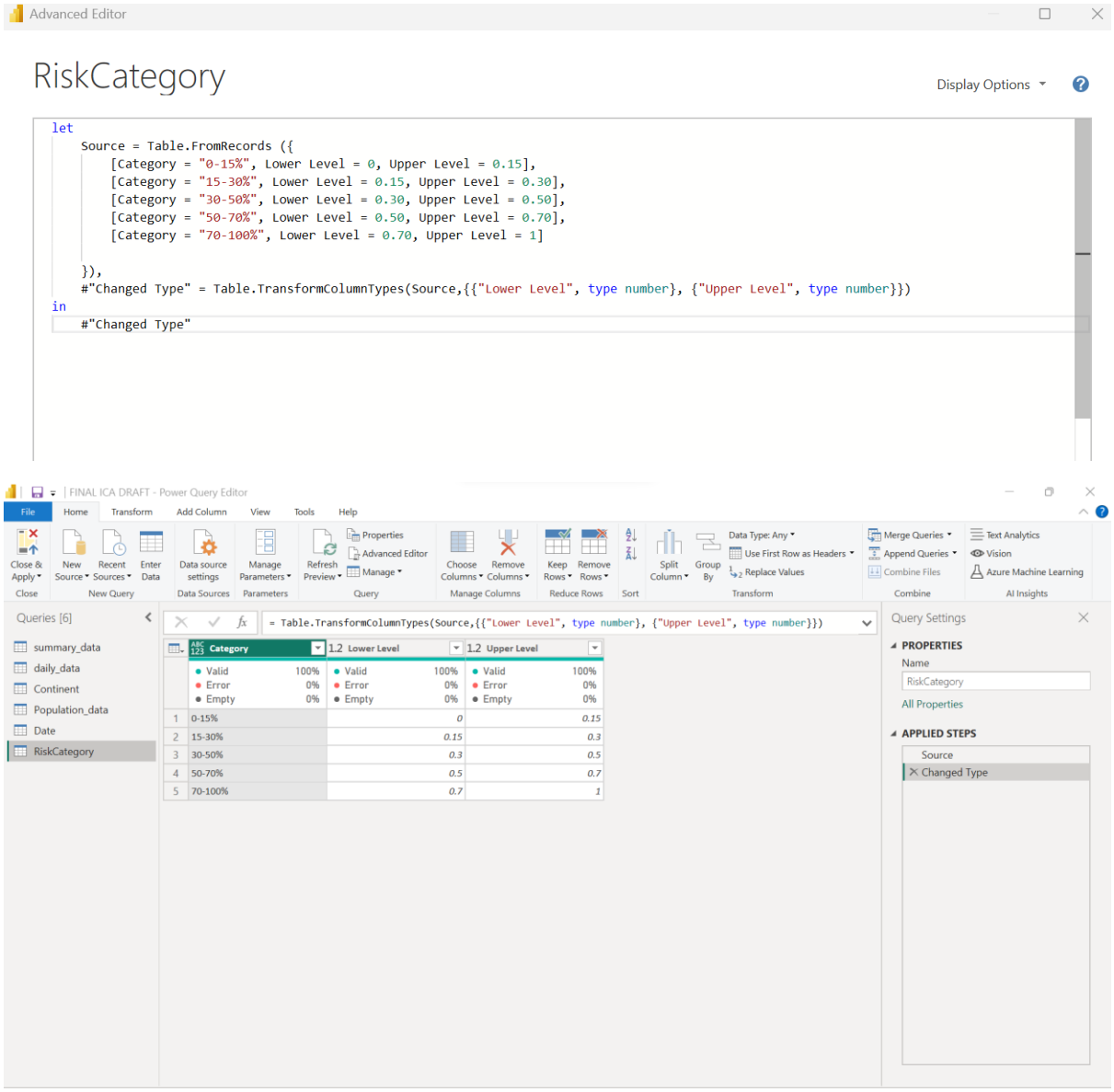



***Figure 7.2: screenshot showing creation of infection risk categories using M Language***

I then converted both columns into percentages. This was done by clicking on the percentage symbol on the **Column Tools** Ribbon.

Next step was to map the risk category table to a risk category column (bin). To do this, I created a new calculated column and called it RiskCategory and used the CALCULATE function to map the infection rate to the risk category. See formula below

```
RiskCategory =
 VAR CatVar=[Infection_rate]
 RETURN
 CALCULATE (VALUES ( RiskCategory[Category]), CatVar>RiskCategory[Lower Level],CatVar<=RiskCategory[Upper Level])
```

**Where**

- VAR = variable which stores the result of the expression which can now be passed as an argument in another measure expression
- CatVar = name of the variable
- [infection_rate] is the expression returning a scalar or table value
- CALCULATE (VALUES ( RiskCategory[Category]), CatVar>RiskCategory[Lower Level],CatVar<=RiskCategory[Upper Level]) calculates the values of the infection rate for each country with respect to their risk category.

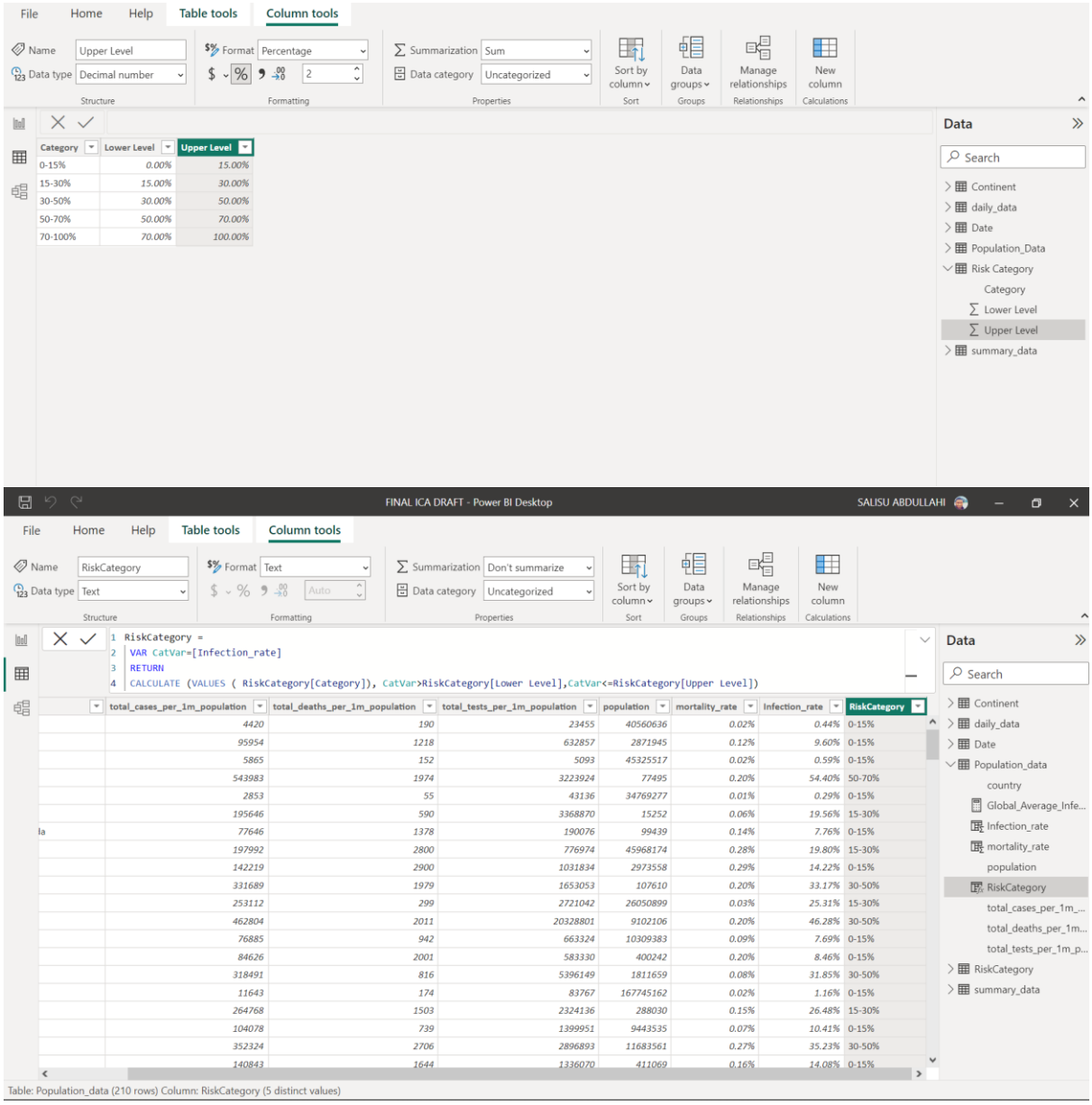
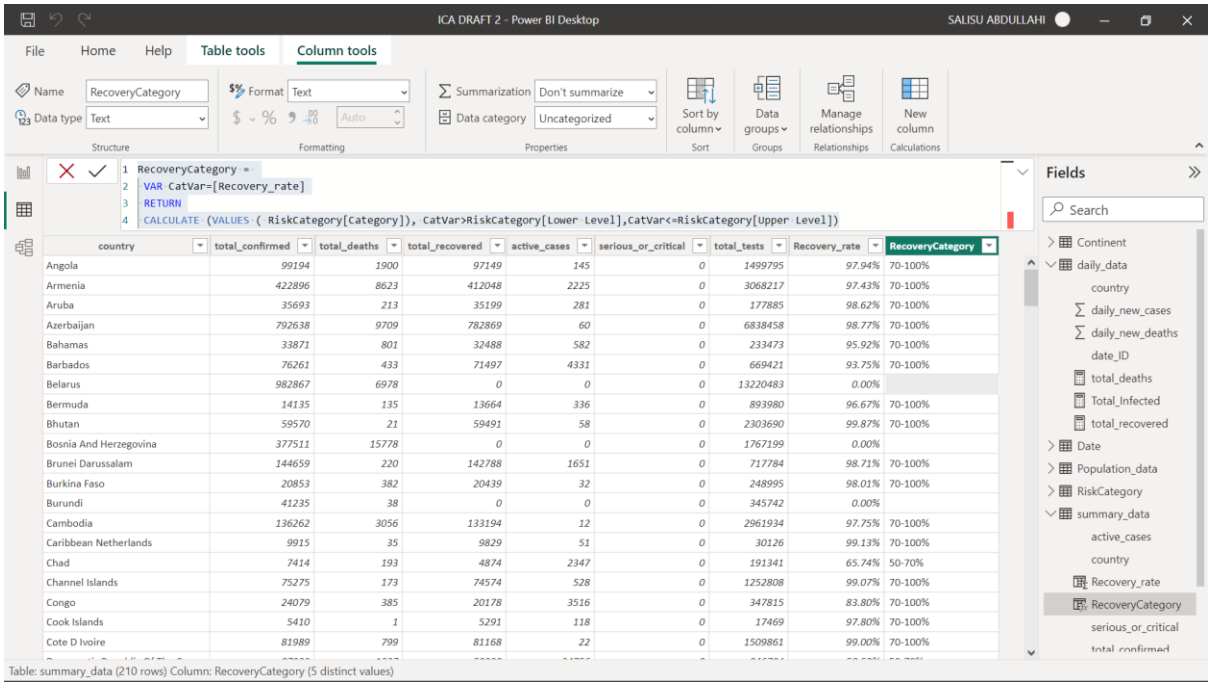



***Figure 7.3: screenshot showing calculation of infection rate into risk categories using M Language***

M. Language was also used to create category for recovery rate as well using the same steps as above but with formula below instead.

RecoveryCategory =

```
VAR CatVar=[Recovery_rate]
RETURN
CALCULATE (VALUES (RiskCategory[Category]), CatVar>RiskCategory[Lower
Level],CatVar<=RiskCategory[Upper Level])
```




| country | total_confirmed | total_deaths | total_recovered | active_cases | serious_or_critical | total_tests | Recovery_rate | RecoveryCategory |
|---|---|---|---|---|---|---|---|---|
| Angola | 99194 | 1900 | 97149 | 145 | 0 | 1499795 | 97.94% | 70-100% |
| Armenia | 422896 | 8623 | 412048 | 2225 | 0 | 3068217 | 97.43% | 70-100% |
| Aruba | 35693 | 213 | 35199 | 281 | 0 | 177885 | 98.62% | 70-100% |
| Azerbaijan | 792638 | 9709 | 782869 | 60 | 0 | 6838458 | 98.77% | 70-100% |
| Bahamas | 33871 | 801 | 32488 | 582 | 0 | 233473 | 95.92% | 70-100% |
| Barbados | 76261 | 433 | 71497 | 4331 | 0 | 669421 | 93.75% | 70-100% |
| Belarus | 982867 | 6978 | 0 | 0 | 0 | 13220483 | 0.00% | |
| Bermuda | 14135 | 135 | 13664 | 336 | 0 | 893980 | 96.67% | 70-100% |
| Bhutan | 59570 | 21 | 59491 | 58 | 0 | 2303690 | 99.87% | 70-100% |
| Bosnia And Herzegovina | 377511 | 15778 | 0 | 0 | 0 | 1767199 | 0.00% | |
| Brunei Darussalam | 144659 | 220 | 142788 | 1651 | 0 | 717784 | 98.71% | 70-100% |
| Burkina Faso | 20853 | 382 | 20439 | 32 | 0 | 248995 | 98.01% | 70-100% |
| Burundi | 41235 | 38 | 0 | 0 | 0 | 345742 | 0.00% | |
| Cambodia | 136262 | 3056 | 133194 | 12 | 0 | 2961934 | 97.75% | 70-100% |
| Caribbean Netherlands | 9915 | 35 | 9829 | 51 | 0 | 30126 | 99.13% | 70-100% |
| Chad | 7414 | 193 | 4874 | 2347 | 0 | 191341 | 65.74% | 50-70% |
| Channel Islands | 75275 | 173 | 74574 | 528 | 0 | 1252808 | 99.07% | 70-100% |
| Congo | 24079 | 385 | 20178 | 3516 | 0 | 347815 | 83.80% | 70-100% |
| Cook Islands | 5410 | 1 | 5291 | 118 | 0 | 17469 | 97.80% | 70-100% |
| Cote D Ivoire | 81989 | 799 | 81168 | 22 | 0 | 1509861 | 99.00% | 70-100% |

***Figure 7.3: screenshot showing calculation of recovery rate into risk categories using M Language***

## Dashboard

My dashboard is made up of 5 pages which analyses and answers the business questions and provide other information. The Pages were titled in chronological order as listed below:

- Global Situation
- Continental Situation
- Country Situation
- Country Specific Situation
- Summary Information

First page provides visual information on covid globally, the second provides a breakdown of covid pandemic across the continents. The third page shows the pandemic effect across the countries under the analysis and the fourth provide information on how COVID has fared in Nigeria and last page provides a summary on the previous dashboards.

## Global Situation Dashboard

This dashboard is made up of 3 cards showing the total infected, total deaths and the current global infection rate respectively. It also contains two drill down timeline zoom charts which visualises confirmed infected cases and confirmed deaths. These charts show the breakdown of covid infected cases by year (2020 to date). Furthermore, these zoom charts allow drill down into any year of choice should you need further analysis of covid cases based on the months of the year and much further into each day of any month. This dashboard is an interactive dashboard as it provides specific details into any selected period as shown below.

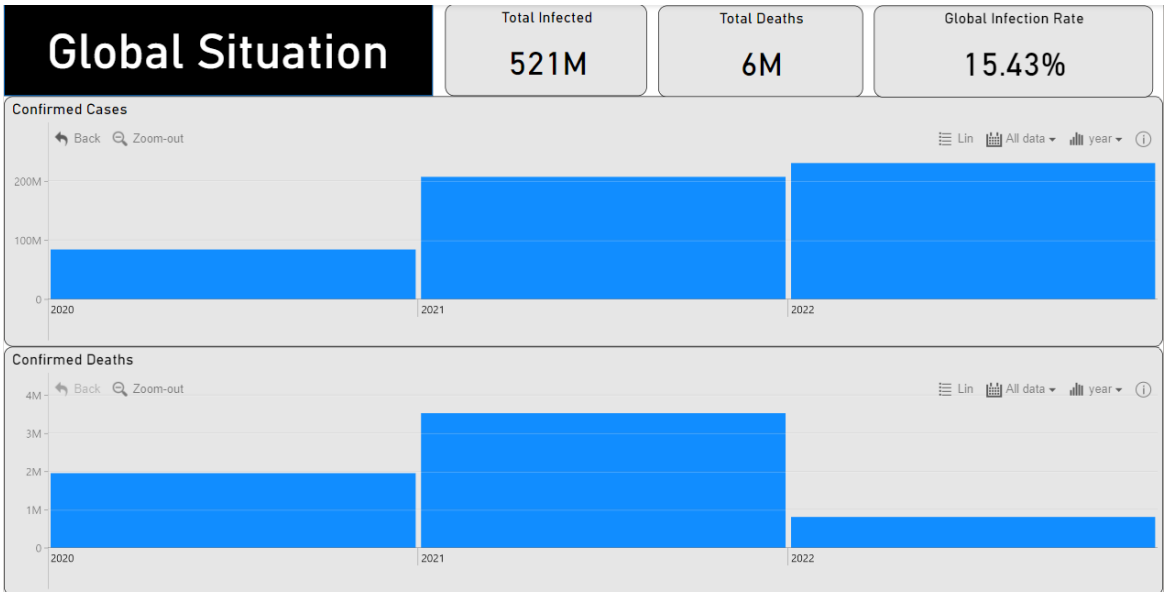



*Figure 8.0: screenshot showing Global situation dashboard of covid by year*

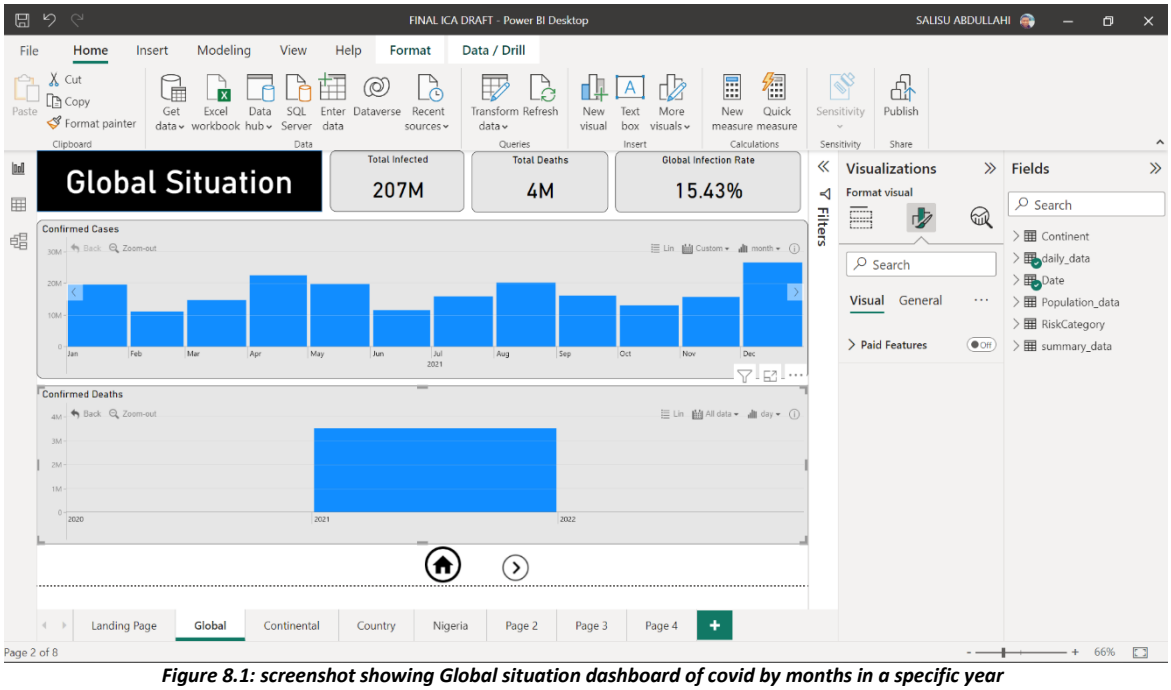



*Figure 8.1: screenshot showing Global situation dashboard of covid by months in a specific year*

## Continent Situation Dashboard

In this dashboard, we have a combination of visuals that show the distribution of infected cases and deaths across the continents of the world while also showing the countries under these continents.

The Visual is made up of a slicer which allows you specify the year and period you want to visualise; an aster plot showing the infected case distribution; a clustered bar chart that shows the distribution of confirmed deaths due to COVID and a multi-row card that shows the countries under any selected continent.

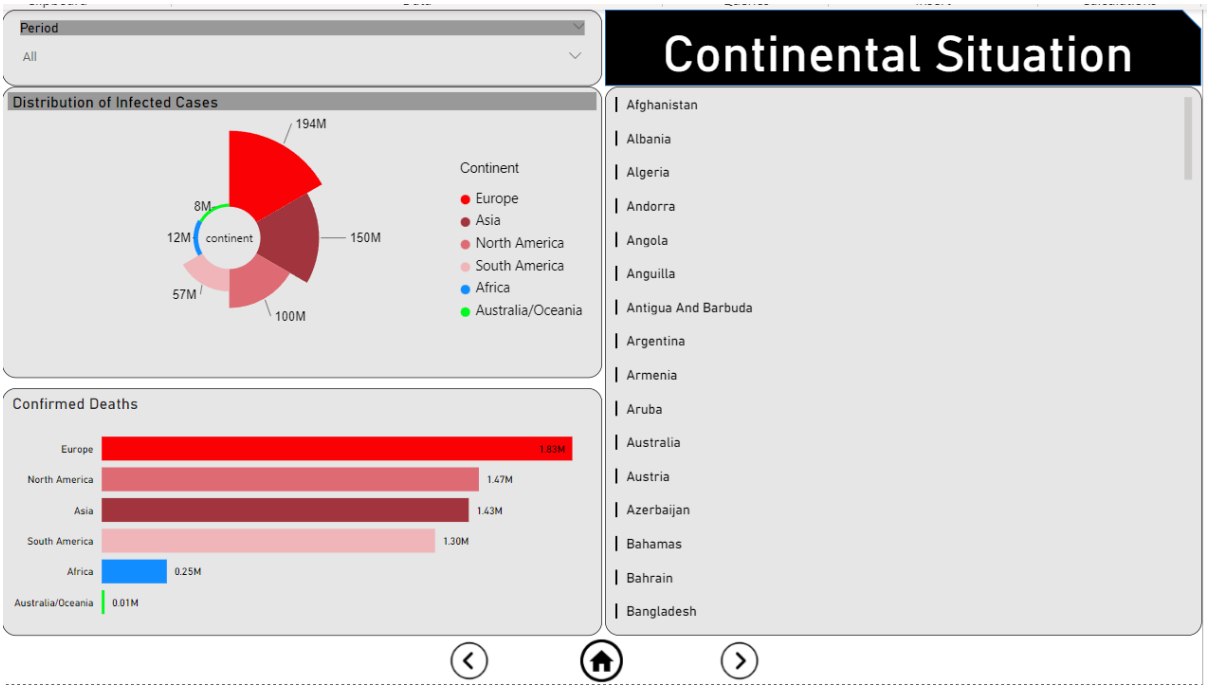



*Figure 9.0: screenshot showing Continental situation dashboard of covid*

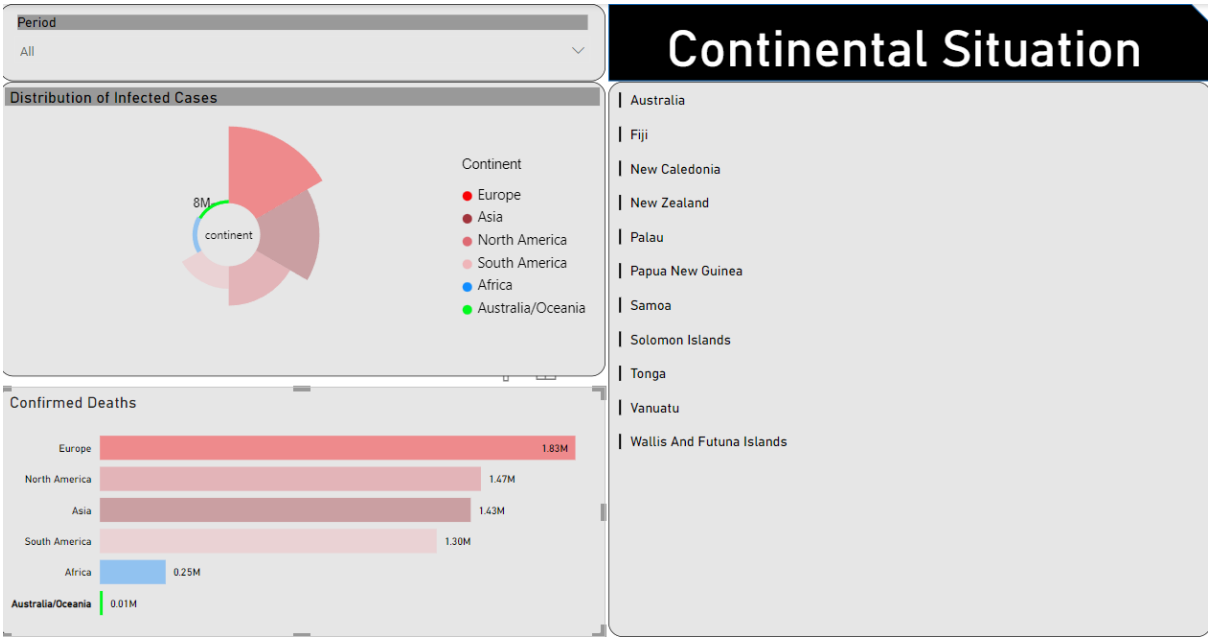



*Figure 9.1: screenshot showing Continental situation dashboard of covid specifically highlighting Australia/Oceania*

## Country Situation Dashboard

Here, I visualised the impact of covid-19 across all the countries under analysis while also providing an infection risk probability for users to advise themselves.

The dashboard contains a table which shows the countries of the world, the total number of infected people, the number of deaths from covid, the infection rate, the covid mortality rate and the covid recovery rate in these countries. The table is also colour-padded to show countries whose infection rate are above the global infection rate average (red) and countries with infection rate lower than the global average(green).

The dashboard also contains a donut chart which visualises the distribution of infection risk across these countries. Countries in each infection risk category are shown under the table when the category is selected.

The last visualisation in this dashboard is a funnel chart which shows the top countries with a high infection rate. The aim of the visualisation is to advise travellers who intend to visit any of the listed countries.

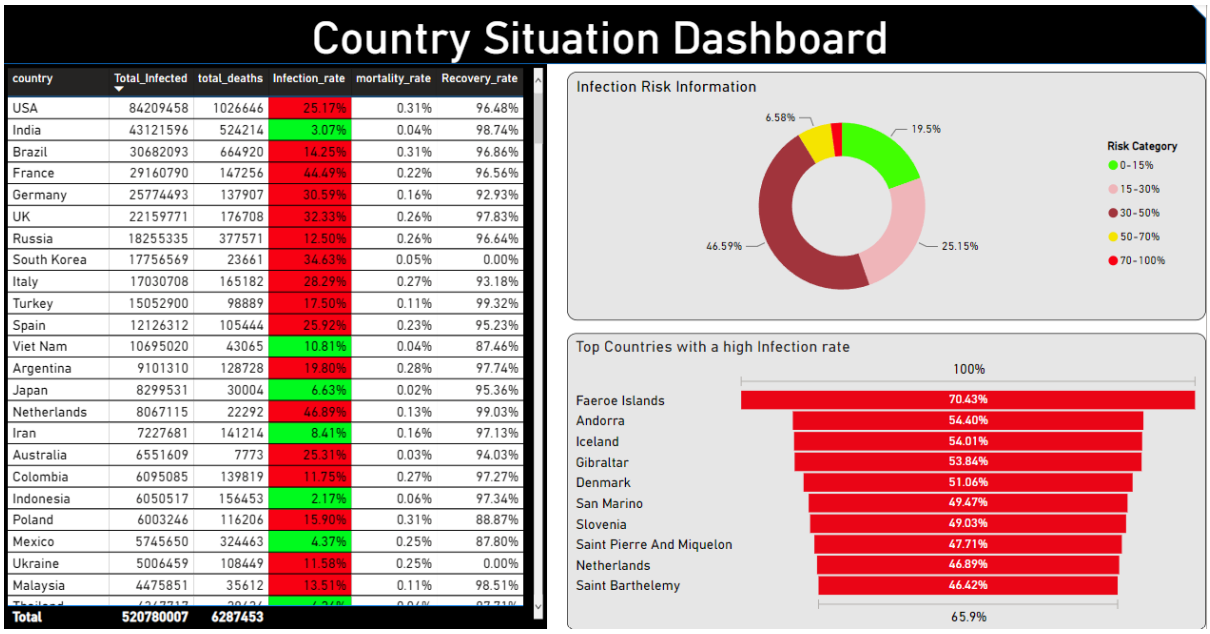



***Figure 10.: screenshot showing Country situation dashboard of covid***

## Country-Specific Situation (Nigeria)

In the country-specific dashboard, I visualised COVID information for Nigeria. This visualisation showed the daily infection statistics and daily death cases since the COVID-19 pandemic started and the influence behind the tendency for the COVID infection to decrease.

This dashboard contains a multi-row card, which provides information regarding the population of Nigeria, the number of covid infection, deaths due to covid and tests for Coronavirus that have been done and the current infection rate in the country

There are also two area charts which shows the daily reported infection cases and the daily confirmed covid deaths across the country respectively since the outbreak of corona virus.

The dashboard also contains a key influencer chart which shows the periods which the average daily number of confirmed infection cases in Nigeria has reduced.

Lastly, for easy navigation, there is a Timeline 2.4.0 slicer which is of advantage should you want information specific to a range of period.

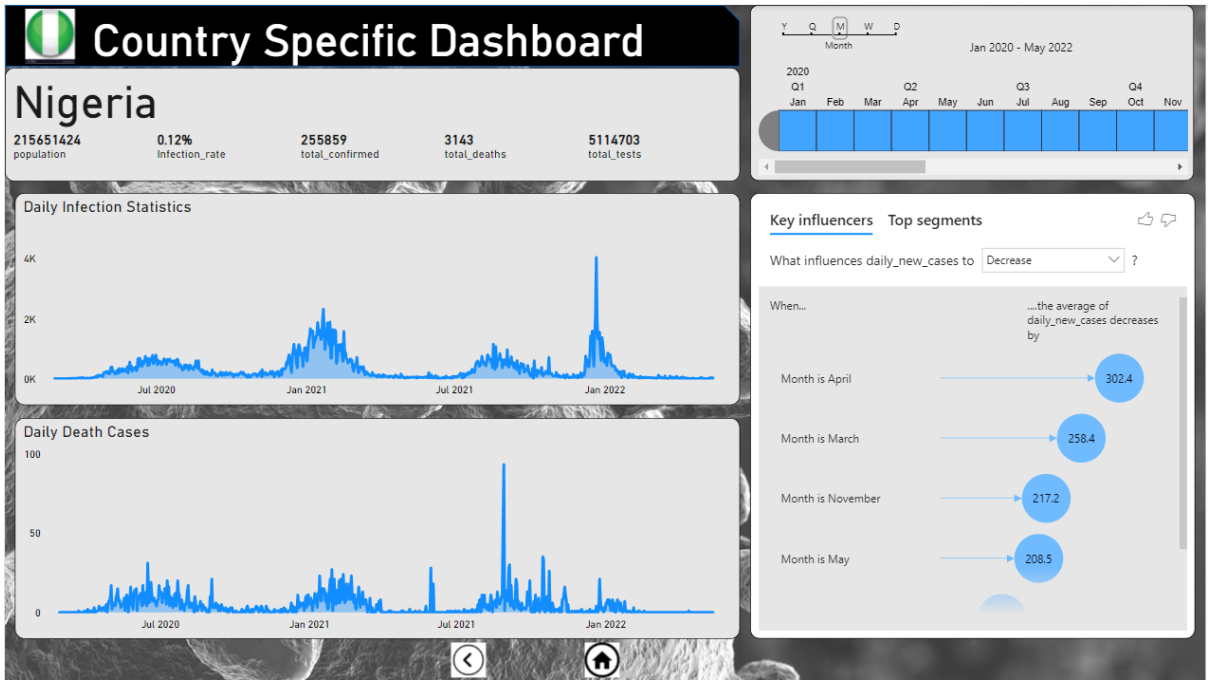



***Figure 11: screenshot showing Nigeria situation dashboard of covid***

## Summary Information

In this dashboard, I visualised the detailed analysis of each of the previous dashboards. This was done using the smart narration visuals which uses artificial intelligence to analyse visualisations and provide insights.

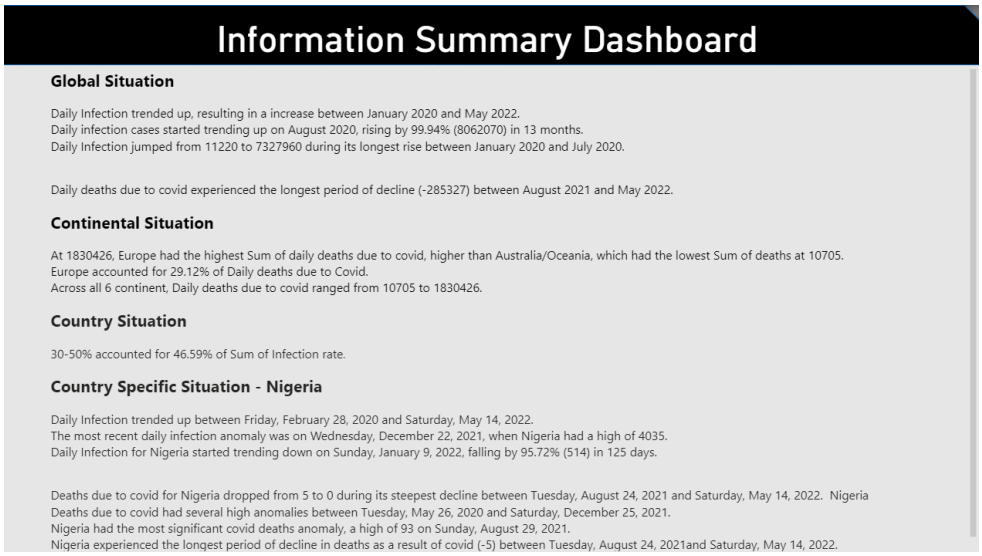



***Figure 12: screenshot showing Summary Information dashboard of covid***

# References

1. Assaker, J. (2022) ‘Covid-19 Global Dataset’. Available at: https://www.kaggle.com/datasets/josephassaker/covid19-global-dataset (Accessed 02 January 2023).

| **Report Section** | **Description** | **Grade your work from 0 to 100** |
|---|---|---|
| Report Structure | The report is well-written, and it contains all the relevant sections | 85 |
| Data Pre-processing and Data Modelling | Many pre-processing steps have been applied. The data model is well-structured | 85 |
| Dax and M language | Both DAX and M Language have been **extensively** used in the report | 80 |
| Dashboard Design | The dashboard contains a variety of charts, including advanced ones not covered in the module. | 90 |
| **Average** | | **Add below the average of the four cells above:**<br><br>**85** |