**Alexandria University**

**Faculty of Computer and Data Science**

**Course Title: Field Training 2023-2024**

**Corona Virus (Covid-19)**

**Field Training Report**

**Demographic analysis Explore how the virus affects different age groups**

Objective of this report: We are going to work on 4 columns which are the three different age ranges which are from 0 to 14, from 15 to 64 and finally from 65 to above and comparing each range of those columns to the number of death, seeing which range had the highest number of death. Now we will show the steps for this report.

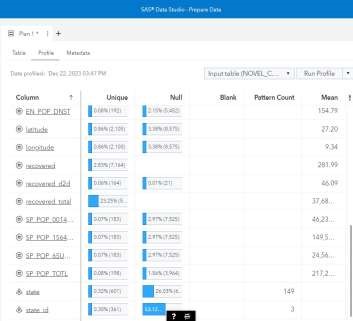
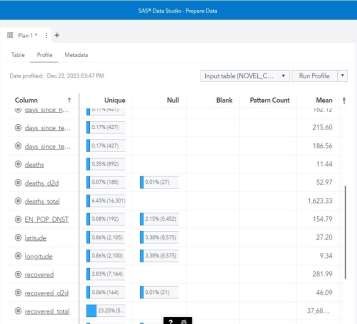
1.

C

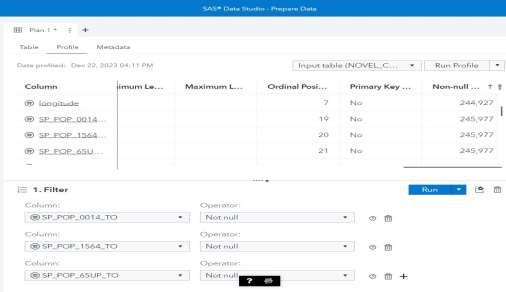
leaning

data

:

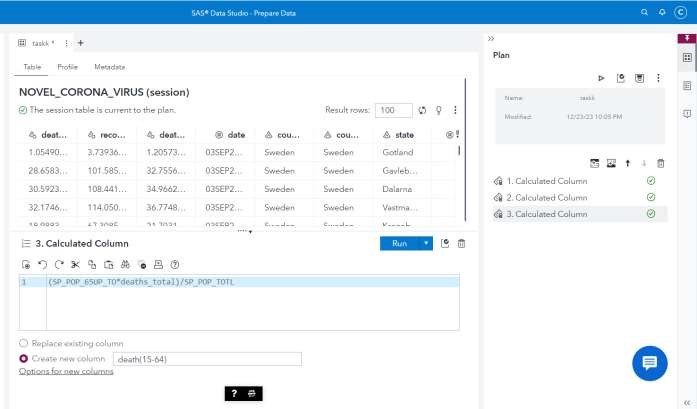


* As we can see total death column has no missing (null) values but as for the three different age range there are empty (null) cells so we need to clean it by removing the empty cell as its percentage is not high so it wont affect our dataset fewer columns will be removed.

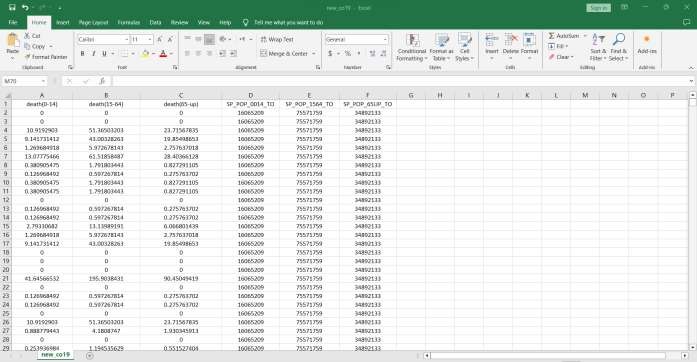


* We used filter to remove the null values from the three age ranges we got.
* Now we will need to create know columns to see how many number of death in each age range as the column (total death) contain the number of total death of the three ranges together.

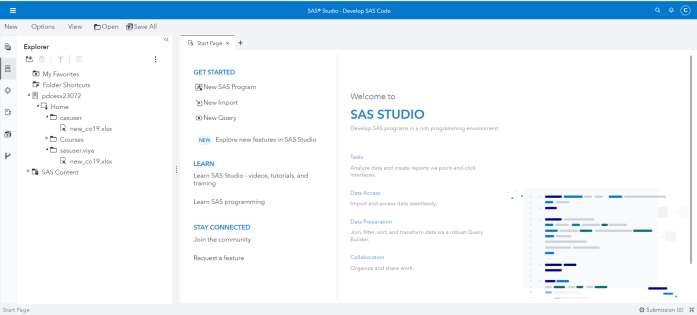
2. Adding new columns.



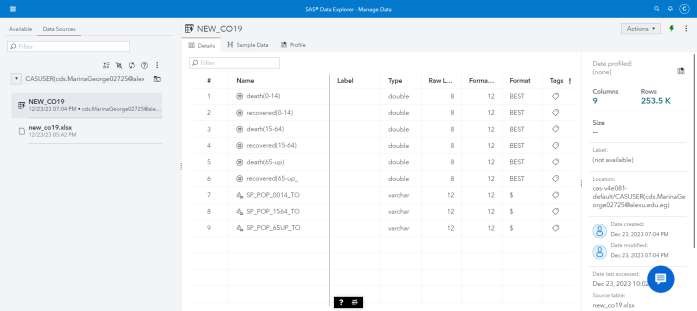
* We calculated three columns so that each range we will know how many death were in this range the equations we used were
  + As for ages from 0-14 total(0-14) \* total death / total population this will give us approximate number of death in the range of 0-14 and store it in a new column.
  + As for ages from 15-64 total(16-64) \* total death / total population this will give us approximate number of death in the range of 16-64 and store it in a new column.
  + As for ages from 65-up total(65-up) \* total death / total population this will give us approximate number of death in the range of 65-UP and store it in a new column.
* Now we got a new table with clean and extra columns so we can download it in our device to upload it on sas as following.



3. Uploading new data set on sas.

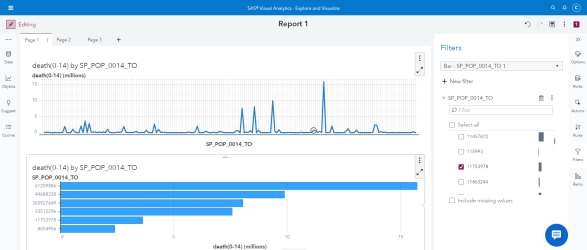


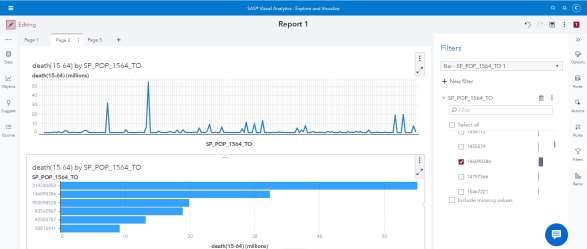
* first in sas develop sas code we will upload our new dataset in sasuser.viya and it will automatically be also uploaded in causer in which we can use in data visualization next steps.

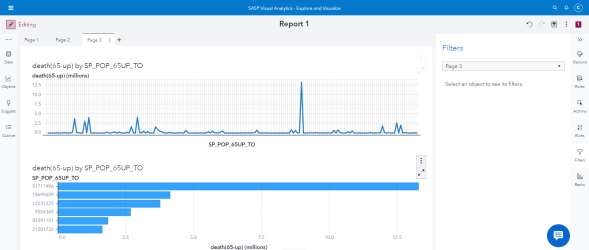


* second we will open our data source from data explorer manage data and import the dataset we just imported.

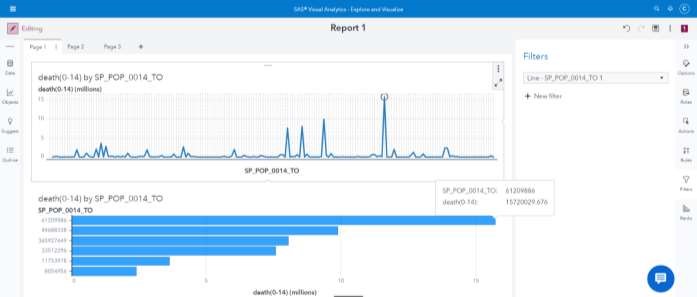
* Now we finished preparing our data by cleaning it and adding columns to do our visual analytics which we will do in the next steps.

1. Sas visual analytics: 

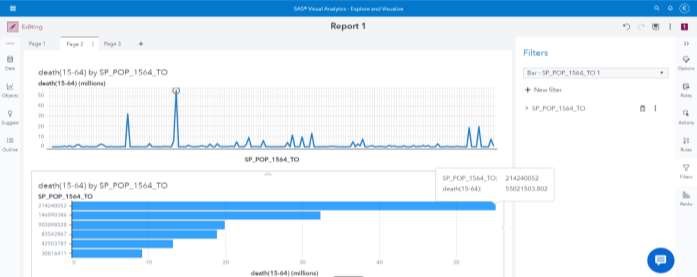
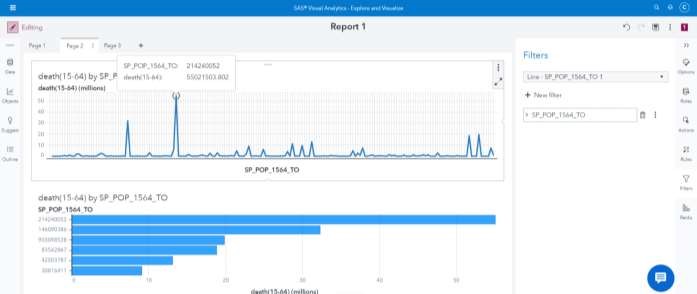




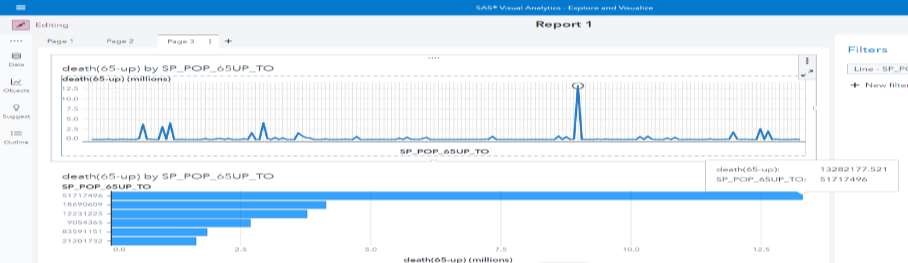
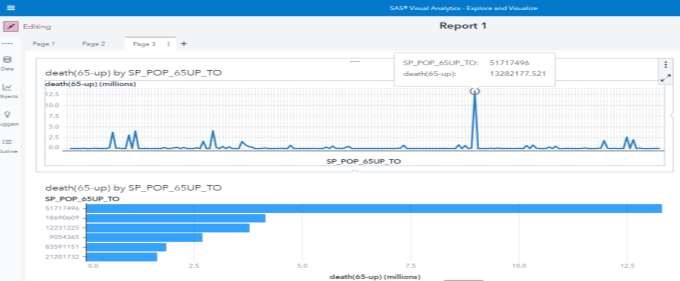
* + First we will do line chart between number of death of each age range and population of each range we will see the numbers at the peeks which mean the highest population with highest number of death and then remove the other numbers (filter) to reduce complexity of the graph and draw the bar plot.
  + By seeing the highest number in line char or bar chart we can determine the highest number of death of the three different age range.



* + as we can see the total number of death for the highest number of group in age range from 0 to 14 is 15,720,029 people who died cause of coronavirus in this group of people in this range.



* + as we can see the total number of death for the highest number of group in age range from 15 to 65 is 55,021,503 people who died cause of coronavirus in this group of people in this range.



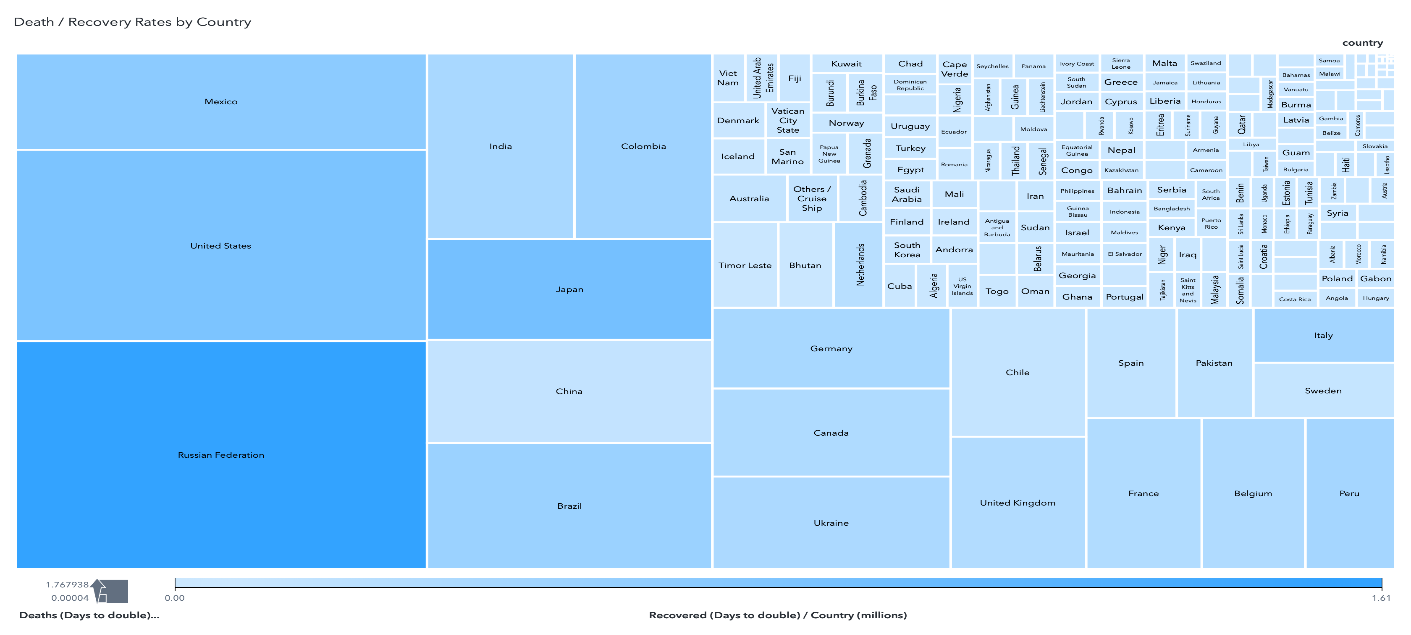
* + as we can see the total number of death for the highest number of group in age range from 65 to above is 13,282,177 people who died cause of coronavirus in this group of people in this range

1. conclusion:

➢ the highest number of death cause of coronavirus was in the age range of 15 to 64 with total number of death 55,021,503. The next patient(victim) probably will be in that range too.

**Death / Recovery Rates by Country**

**Tree Map**

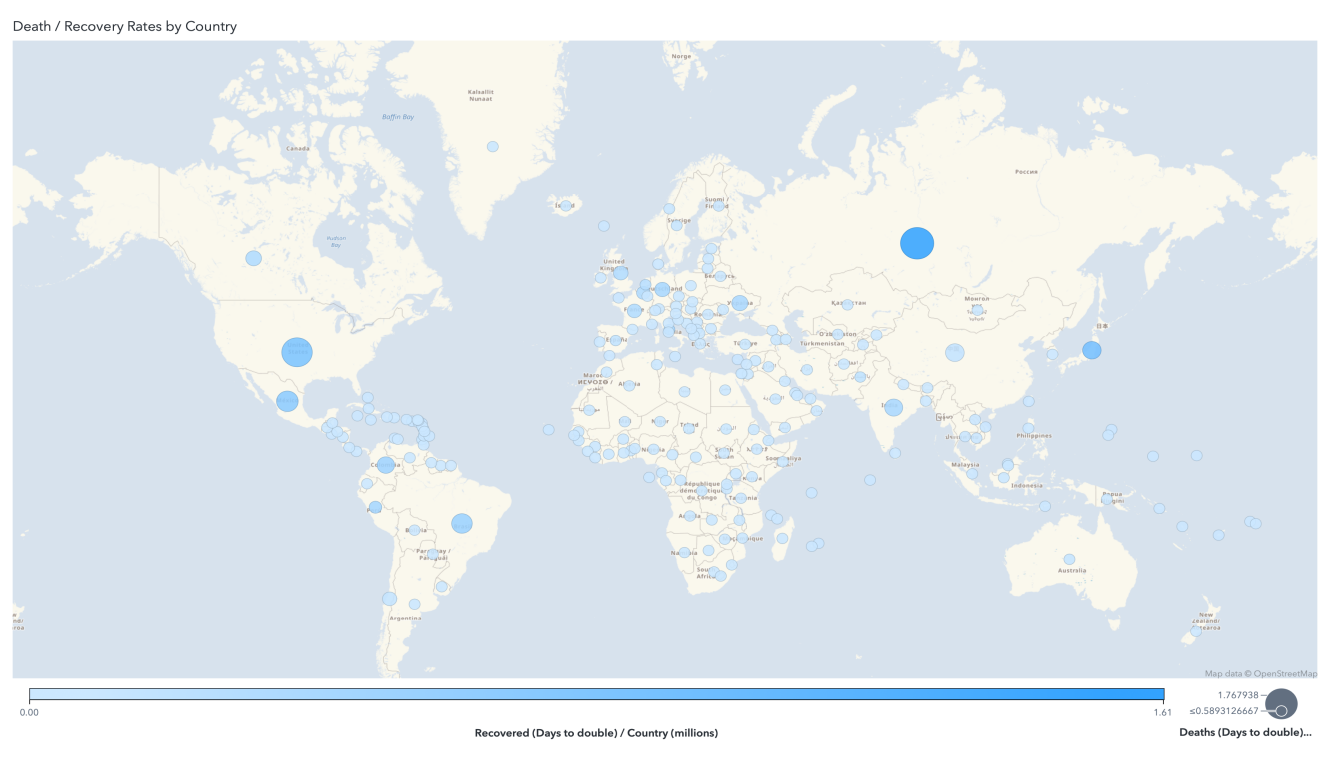
****

In this tree map, 3 variables of coronavirus dataset are added: Country, Recovered (days to double ) / country , and Death ( days to double ) / country.

Recovery rates are measured by the darkness of the blue color, the darker the blue color, the higher recovery rates the country has.

Death rates are measured by the size of the rectangle, the larger the rectangle, the higher death rates the country has.

**Geo Map**

****

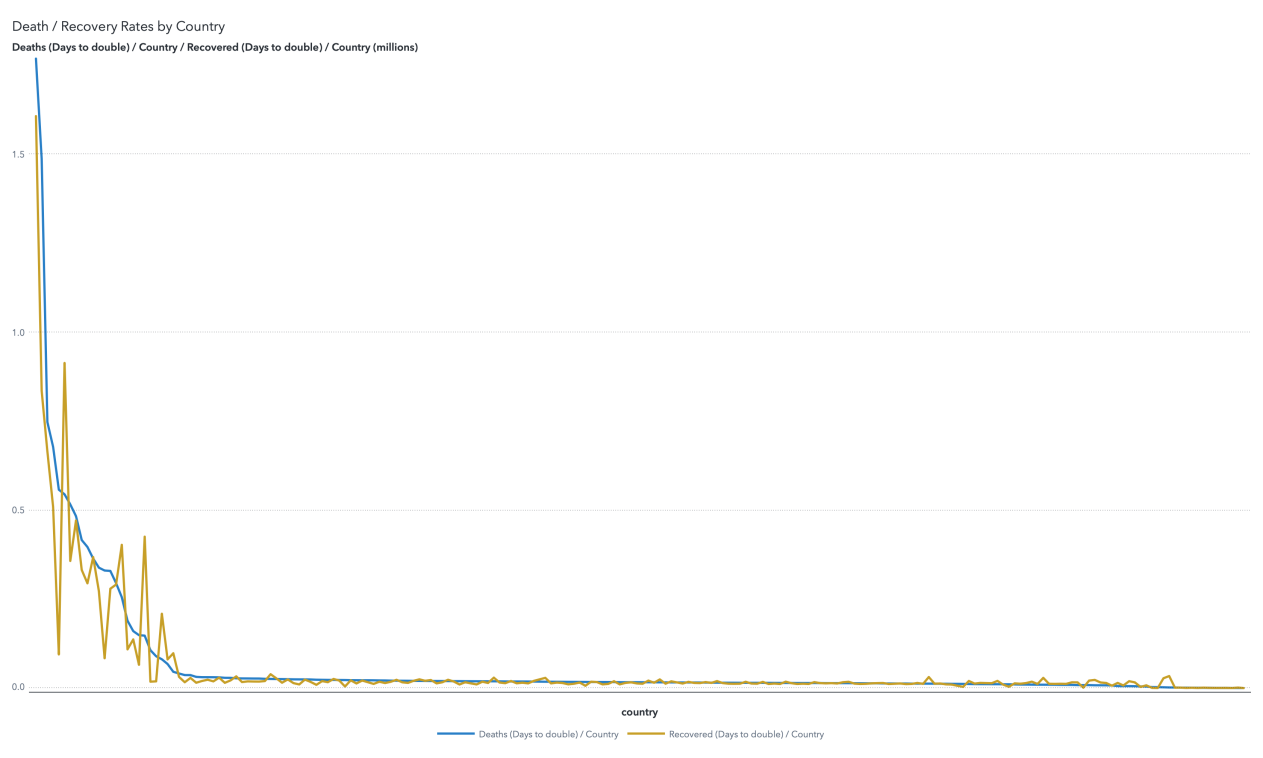
In this geo map, 3 variables of coronavirus dataset are added: Country, Recovered (days to double) / country , and Death ( days to double ) / country.

Here, there are circles located on each of the 212 country.

Recovery rates are measured by the darkness of the blue color of the circle, the darker the blue color, the higher recovery rates the country the circle is located on has.

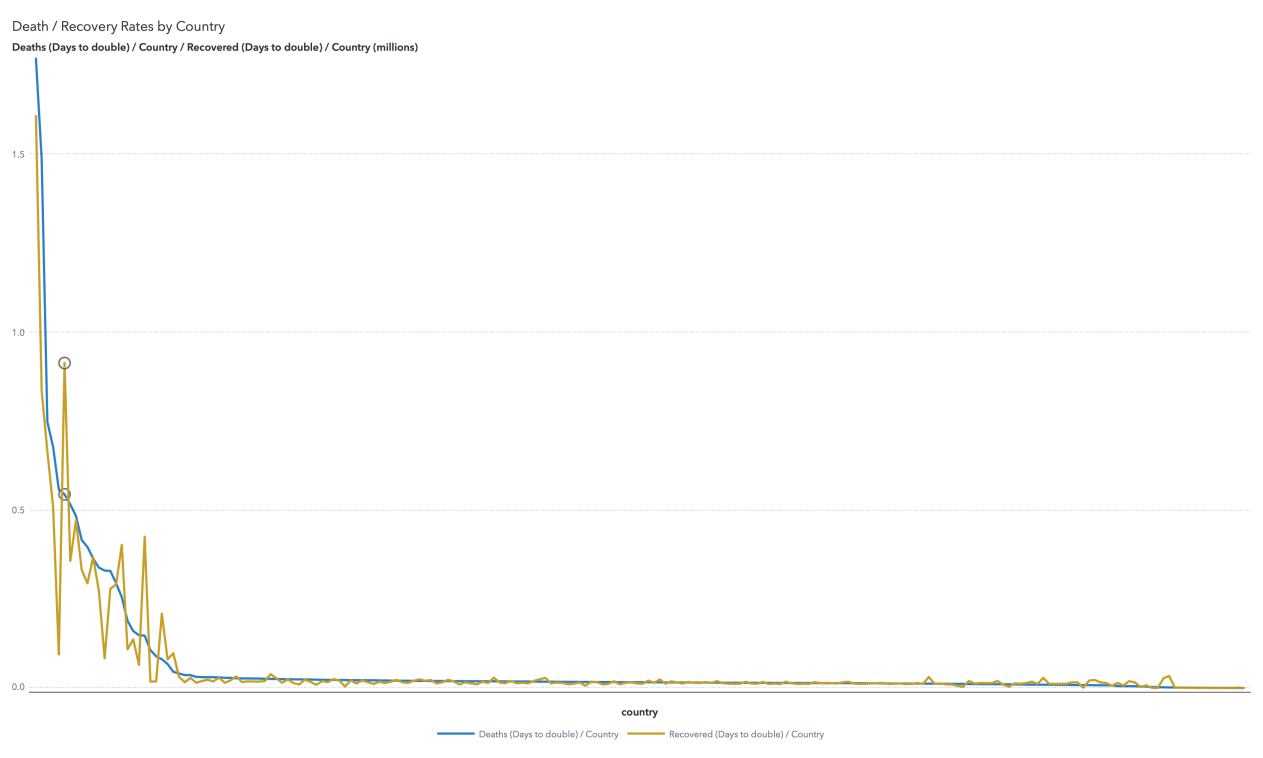
Death rates are measured by the size of the circle, the larger the circle, the higher death rates the country has.

**Line Chart**

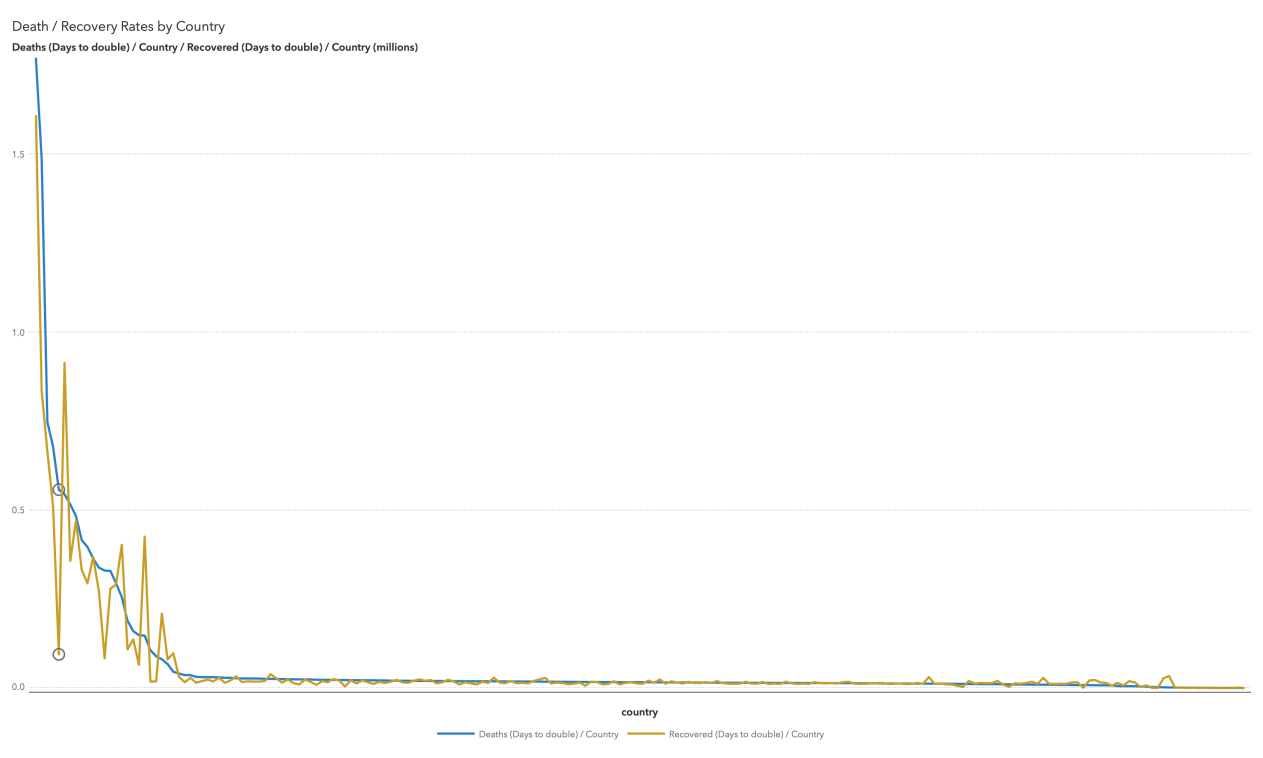
****

The country with the best healthcare system is the country with the greatest gap between its recoveries and deaths rate (recoveries being greater), and the country with the worst healthcare system is the country with the greatest gap between its recoveries and deaths rate (deaths being greater).

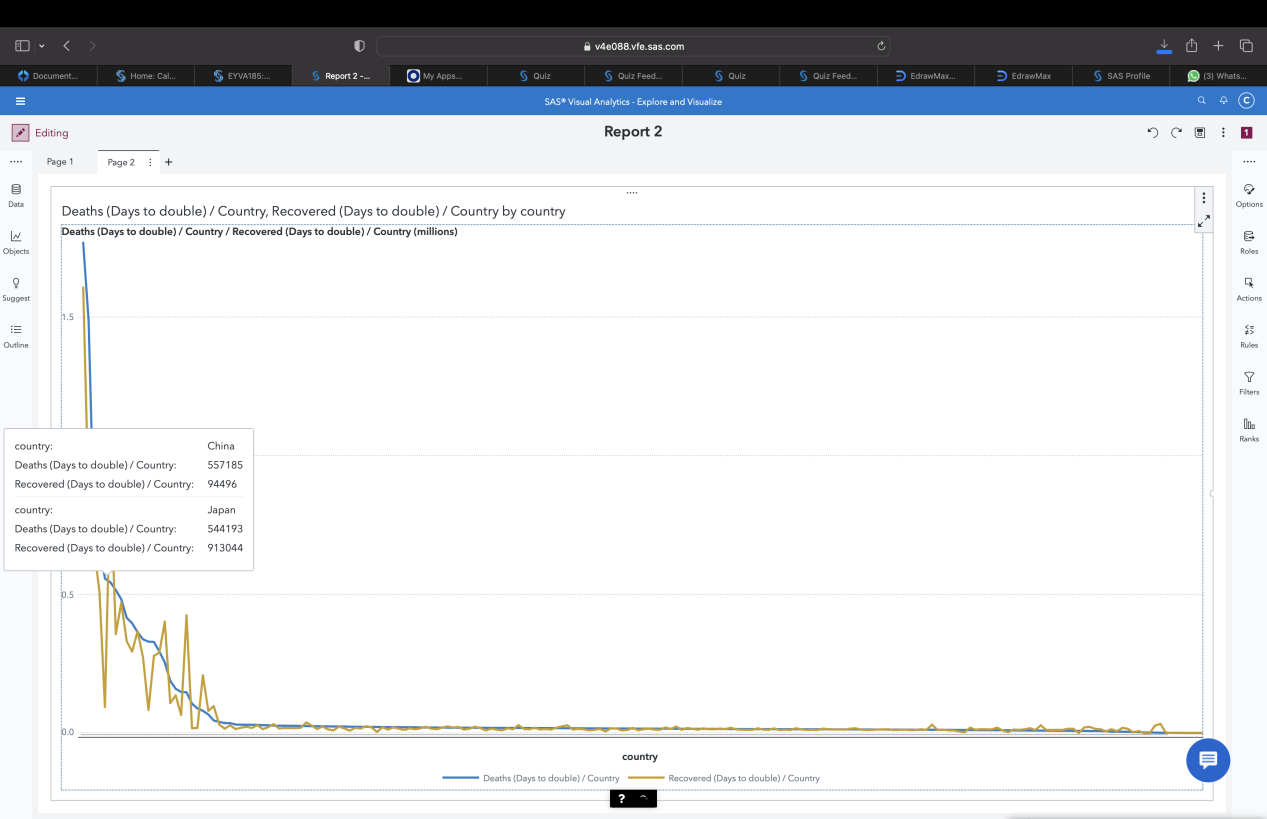
This line chart shows the increase in deaths and recoveries rate.

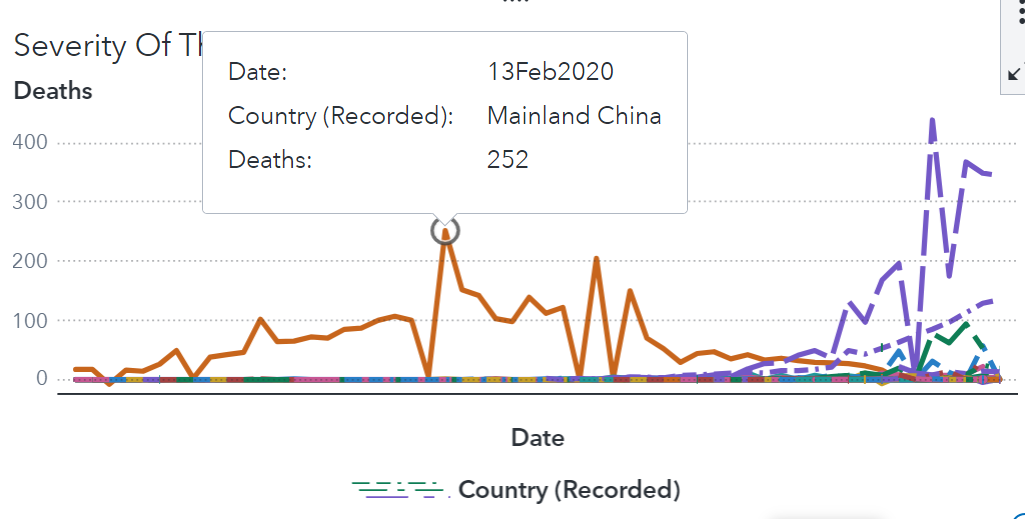


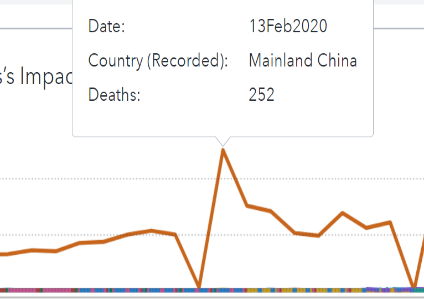
Japan is the country located at this point. As we can see, the two lines has a big gap at this point, with the recovery rate being much higher than the death rate, which means that Japan is the country with the best healthcare system.



China is the country located at this point. As we can see, the two lines has the biggest gap at this point, with the death rate being much higher than the recovery rate, which means that China is the country with the worst healthcare system.





At the beginning the deaths were relatively low for all countries but one country around the middle of the timeline (Mainland, China). It started on January 22th with relatively low death rates until

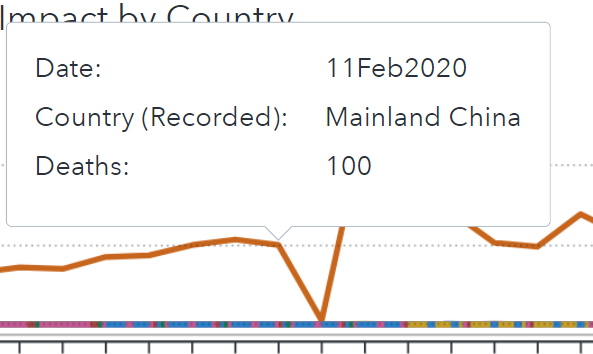
February 13th was the day Mainland, China witnessed one of its worst outbreaks and spikes in its records and other countries’ too.

Even though death rates were relatively high throughout the month ranging from (70-107)

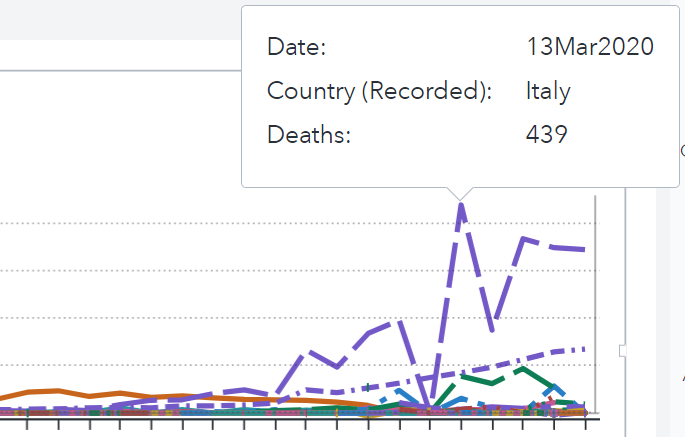
It never spiked as much as did on the 13th of the month.

It was unexpected because just the day before on February 12th the death rates have gone down by nearly 95%.

It has gone through multiple other spikes throughout the year, but it was never as bad as the one on February 13th of 2020.



Most countries had consistent average death rates throughout the year, which were quite low compared to China, until March 7th when most countries started experiencing some disturbances too, specifically, Italy .It experienced the most death rates and the worst spikes of all countries till the end of March.



Same slope as China’s (From relatively high to extremely low then to extremely high) but with much higher death rates. The month ended with it still being on top at 345 deaths.

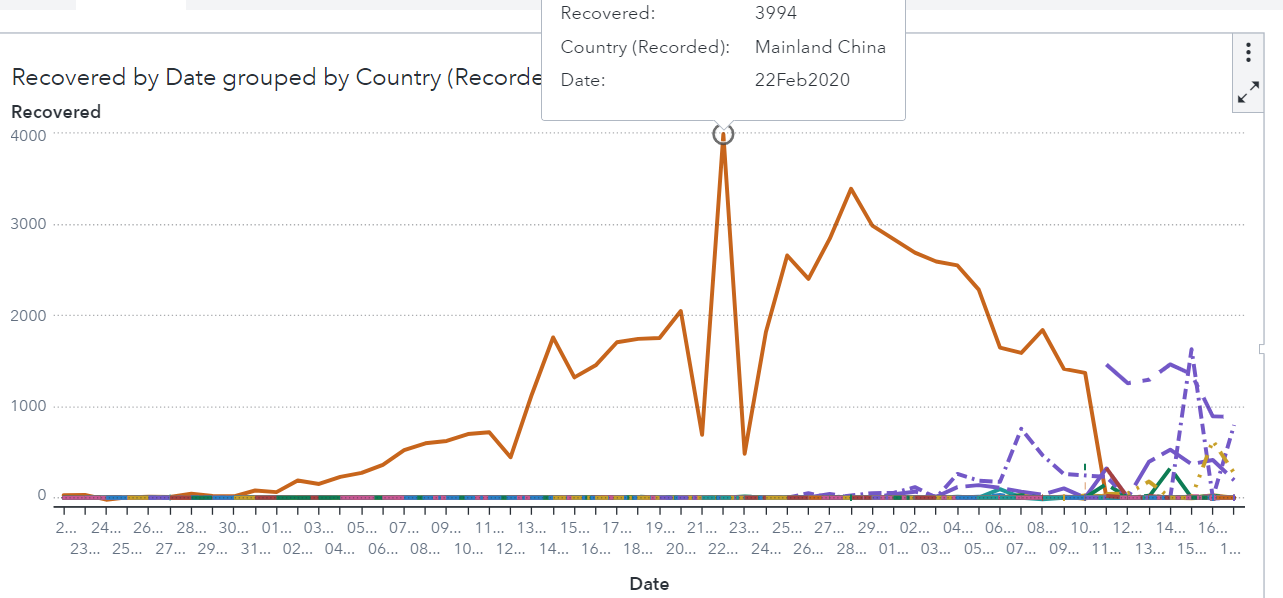
We could see a pattern in both graphs and it is that after each fast drop in death rates

(When countries were on locked down and everyone was quarantined to slow down the virus spread rate by lowering human contact levels)

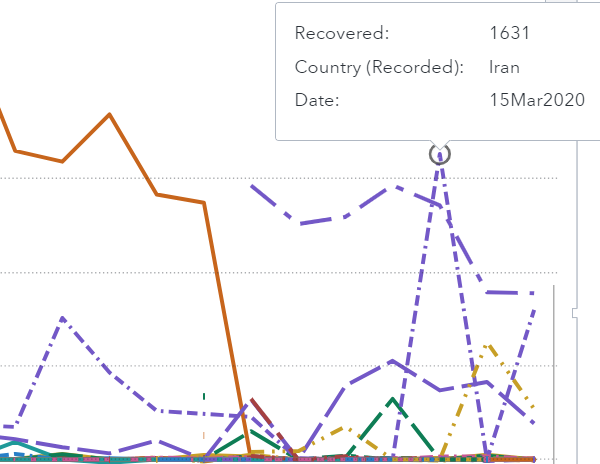
The slope always increased as fast as it dropped.

If we could conclude one thing it would be that human contact had little to no correlation to virus spread rate, further studies need to be made in other causation areas to figure out what caused these spikes.

The virus has affected different countries differently. Some countries have had a relatively low number of deaths throughout the timeline, while others have experienced a significant spike in deaths.



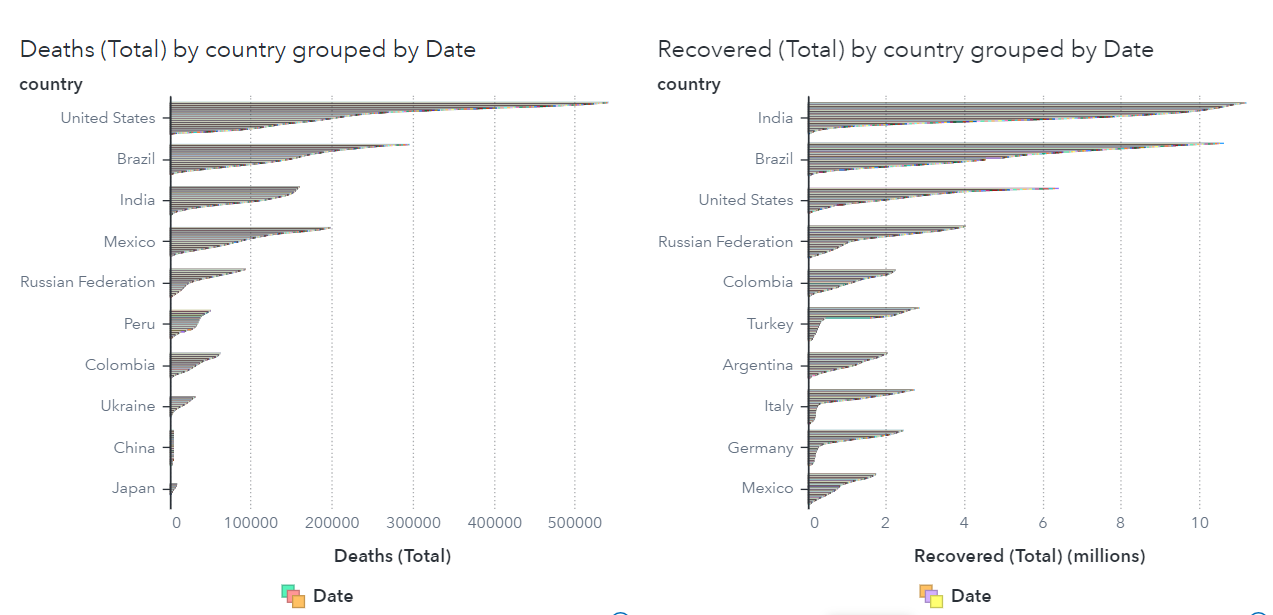
As for as for the recovery rates, they were low throughout the whole period except for one country, the one with the most death rates had the highest recovery rates too, (Mainland, China) until March 10th when China’s recovery rates slowed down extremely and had the same low rate as the rest of the countries.



There was another country that witnessed some high recovery rate spikes at the end of the period and it was, Iran.

Iran’s recovery rate fluctuated most of March ranging from extremely high to extremely low, multiple times in only a couple of days.

But on March 7th and 15th, Iran’s recovery rates reached their all-time high with 756 patients recovered on the 7th and 1631 on the 15th of the month.



Finally, when comparing total deaths and total recoveries we can conclude that even though china was one of the first country exposed to the virus, surprisingly, it wasn’t the most affected (USA, Brazil, and India ranked highest)

USA was ranked highest in total deaths with a big gap in its amount of deaths relative to the countries ranked right blow it, which showed that people in the USA weren’t as cautious as other countries when dealing with covid, not at the virus’ origin and neither half way through.

We can also observe that the top countries with total deaths ranked top countries in total recovery too and neither of them was china nor Italy.

India was the only country that had a consistent average of total recoveries throughout the period, that meant that there were a lot of people affected but not all of them were dead, that is why it is ranked third and not higher on the total deaths pedestal, which means that most people who were affected by the disease died in the USA.

**Regional Analysis:**

Here is a regional analysis to analyze the spread of the virus across different countries and to determine the best healthcare protocols provided by different countries.

1. **Geo coordinate map:**

In this map I wanted to know what countries’ death propagation rate with respect to confirmed cases rate.

Steps to build up the graph:

**Data selection:**

A screenshot of a computer

Description automatically generatedColumns selected*: country, {Confirmed (days to double)/country and Death(days to double)/country }* both aggregation rule is Sum

**Roles setting:**

The *country* column is used as Geography.

Size of Bubble is determined by the *Confirmed (days to double)/country*

Color of the bubble is determined by

the *Death(days to double)/country*

**Applying filters:**

A screenshot of a computer

Description automatically generated3 filters applied on the 3 columns:

1-check that ***country*** isn’t

missing valued country.

2-set the ***Confirmed (days to double)/country*** value more than 0.

3-set the ***Death (days to double)/country*** value more than 0.

**Insights we have:**

The *Confirmed (days to double)/country* determines the time it takes for the cumulative number of confirmed cases to be doubled. So, countries with lower *Confirmed (days to double)/country* value (**Smaller bubble size**) indicates a faster growth rate in confirmed cases while countries with a higher *Confirmed (days to double)/country* value (**Larger bubble size**) indicates slower growth rate in confirmed cases.

The *Death(days to double)/country* determines the time it takes for the number of deaths in a specific country to double. So, countries with a lower *Death(days to double)/country* value (**Light colored bubble**) indicates faster growth of fatalities while countries with a higher *Death(days to double)/country* values(**Bold colored bubble**) indicates more effective mitigation measures.

A screenshot of a computer

Description automatically generated

From the graph we can inspect: **Russian Federation** have the highest propagation rate and slowest growth rate of fatalities.

A screenshot of a computer

Description automatically generated**List table:**

While **Saint Martin** has the lowest propagation rate but has the highest growth rate of all fatalities.

In this list table we will list all the recovery and death percentage for all countries that have recorded cases.

And the difference between the death and recovery rates will indicate the healthcare level in different countries.

Steps to build the List table:

A close up of a white background

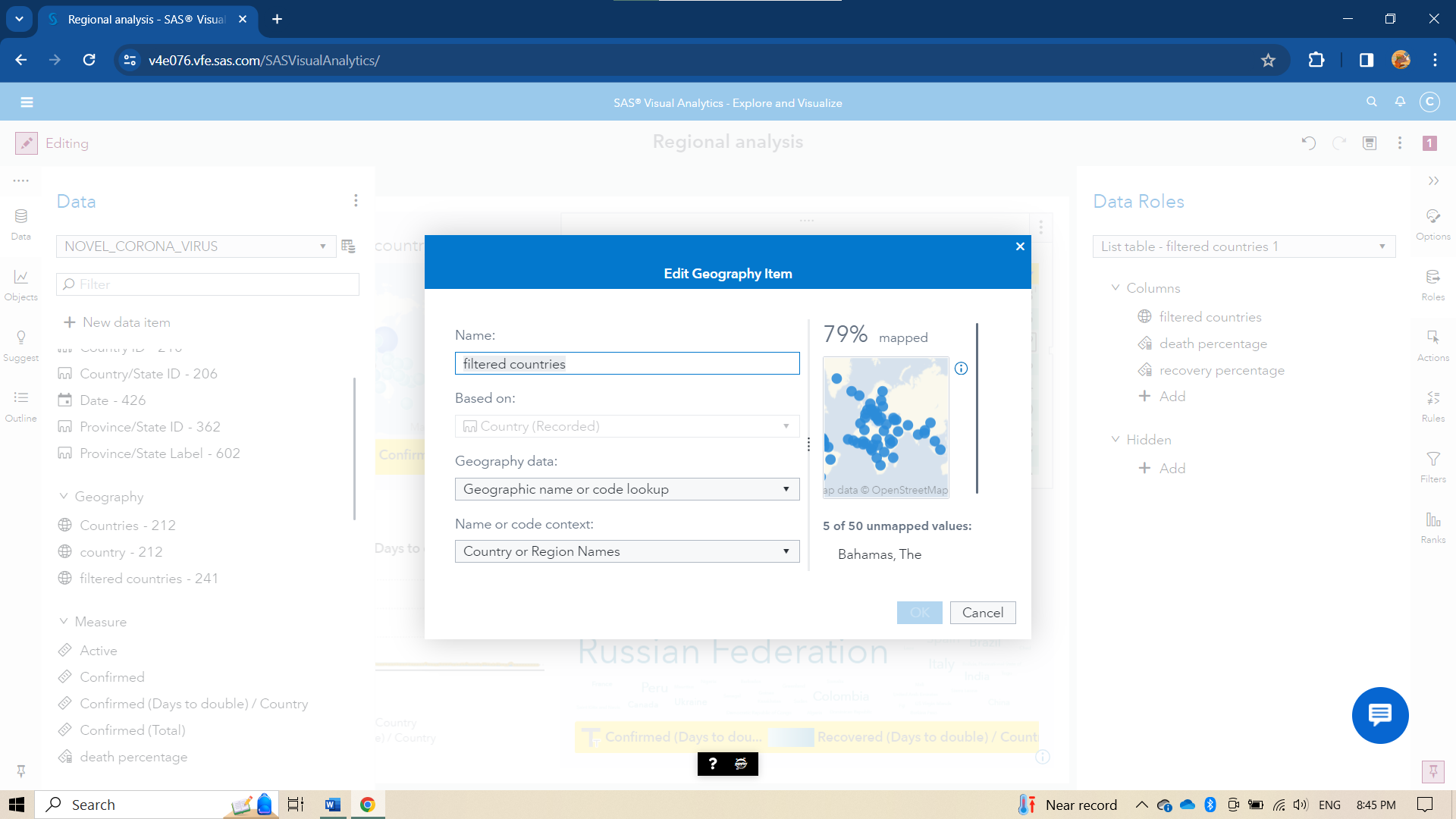
Description automatically generated**Data preparation:**

New calculated items are calculated:

A close-up of a sign

Description automatically generated-*death percentage*

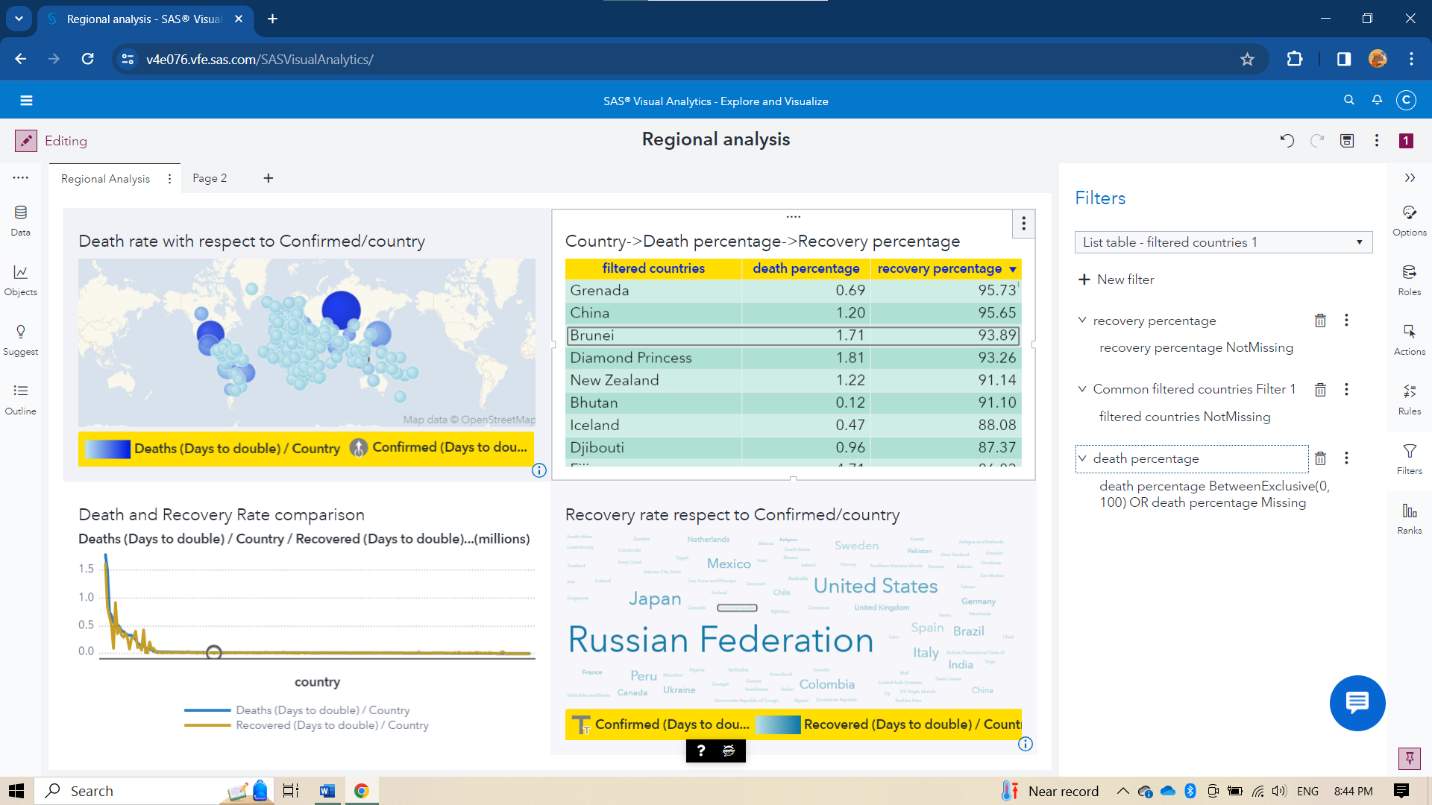
-*recovery percentage*



And a geography item named *filtered countries* depends on *Country (Recorded)* is added to include only countries we have a recorded cases in.

**Data Selection:**

-3 columns are selected

*-death percentage*

*-recovery percentage*

*-filtered countries*

**Applying filters:**

-3 filters are applied

-For *filtered countries* to put only not missing values.

*-recovery percentage* to put only not missing values.

*-death percentage* to put only not missing values.

**Insights we have:**

All countries are listed with the *recovery percentage & death percentage*.

It’s power is spotted on linking it with other graphs in report as will be shown.

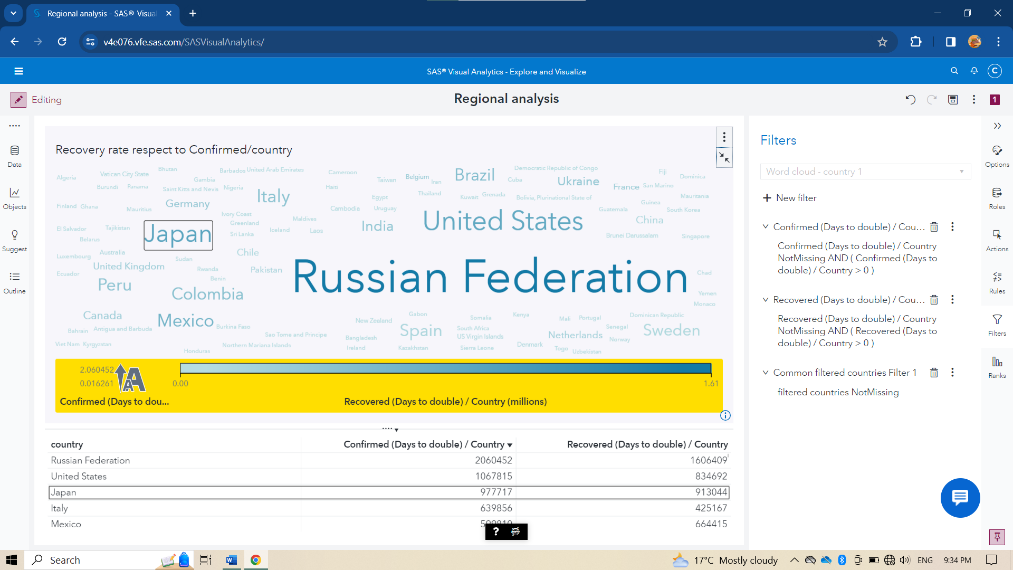
**Word cloud:**

In this word cloud I wanted to know what countries’ recovery propagation rate with respect to confirmed cases rate.

Steps to build the Word cloud:

**Data selection:**

3 columns are selected: *country, { Confirmed(days to double)/country and Recovery(days to double)/country } aggregated by sum.*

**Applying filters:**

3 filters are applied:

For *country* to put only not missing valued.

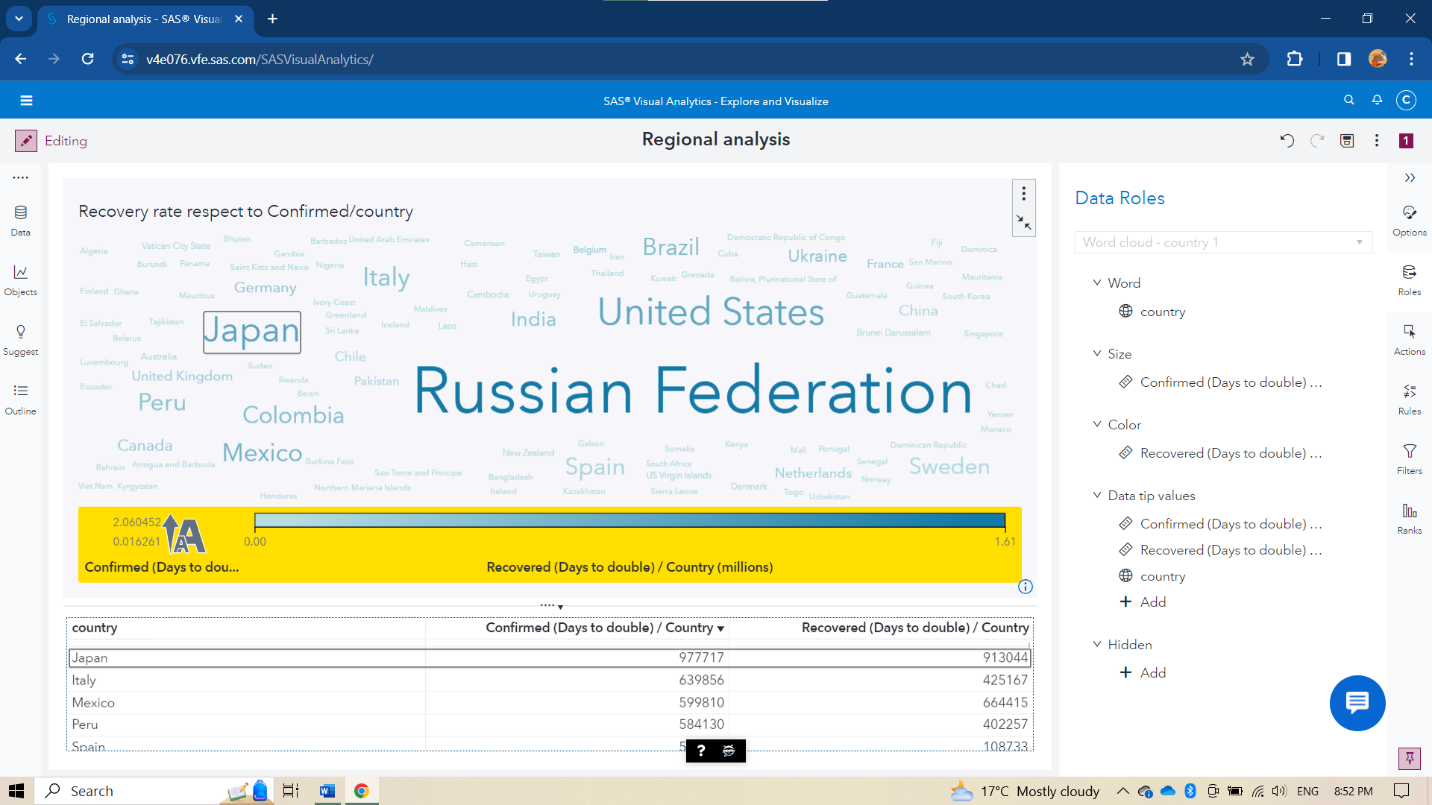
For *Confirmed (days to double)/country* to include only values more than 0 and not missing.

For *Recovery (days to double)/country* to include only values more than 0 and not missing.

**Insights we have:**

The *Confirmed (days to double)/country* determines the time it takes for the cumulative number of confirmed cases to be doubled. So, countries with lower *Confirmed (days to double)/country* value(**Smaller word size**) indicates a faster growth rate in confirmed cases while countries with a higher *Confirmed(days to double)/country* value(**Larger bubble size**) indicates slower growth rate in confirmed cases.

The *Recovered (days to double)/country* determines the time it takes for the number of recoveries in a specific country to double. So, countries with a lower *Recovery (days to double)/country* value (**Light colored word**) indicates faster rate of recovery while countries with a higher *Recovery (days to double)/country* values (**Bold colored word**) indicates challenges in healthcare resources.

Japan indicates with it’s number of *Confirmed (days to double)/country* that the number of confirmed to be doubled is not as high as Russian Federation and not as low as China and The *Recovered(days to double)/country* indicates that recovery rate is high with respect to propagation rate.

**By the help of Line chart made:**

Steps to build the Line chart:

* **Data selection:**

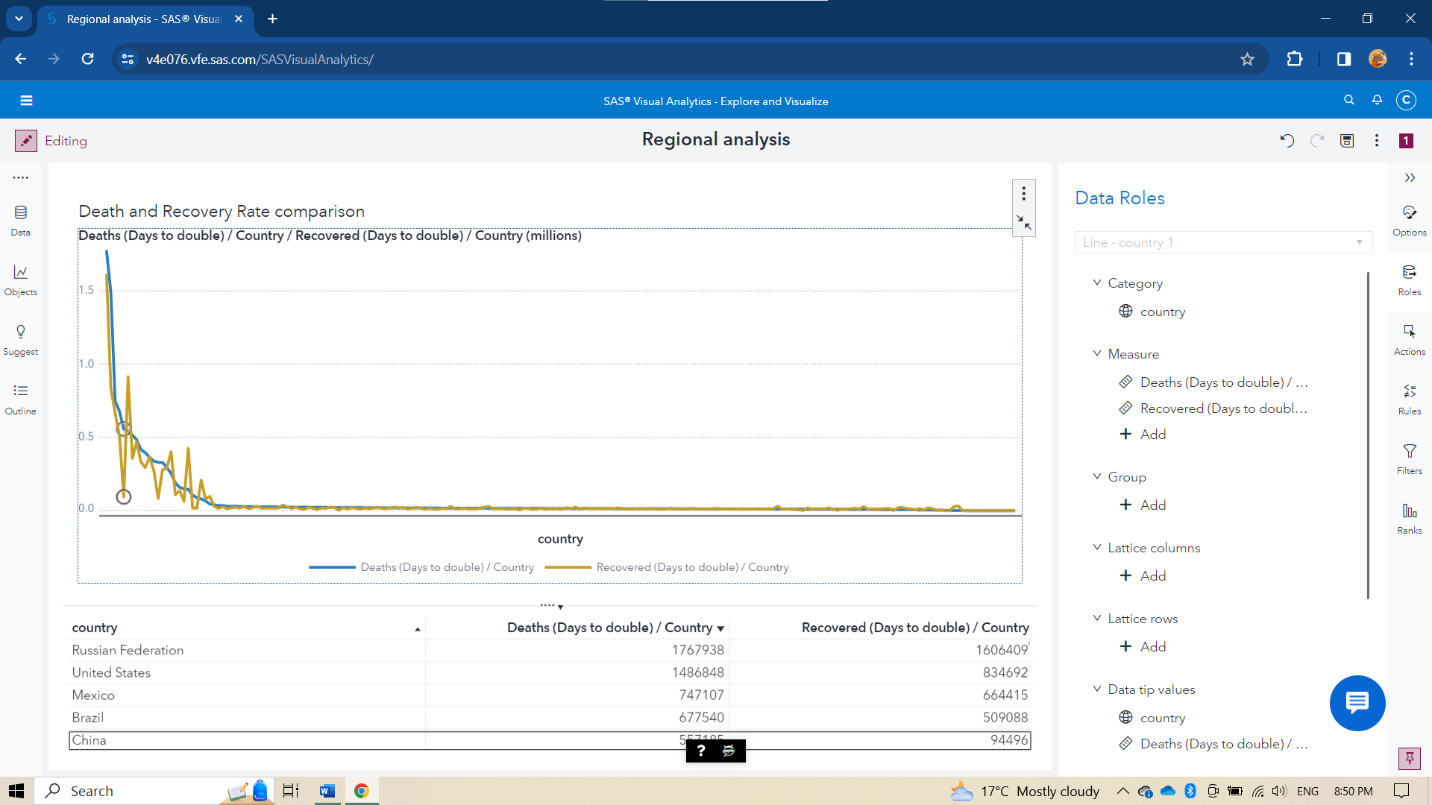
3 columns selected

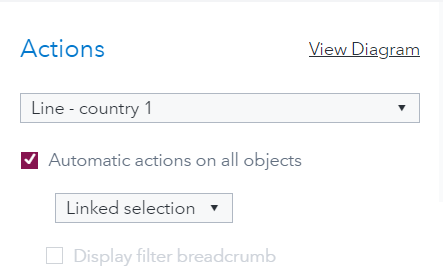
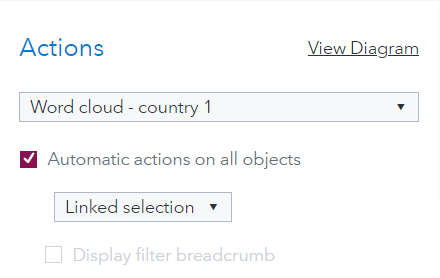
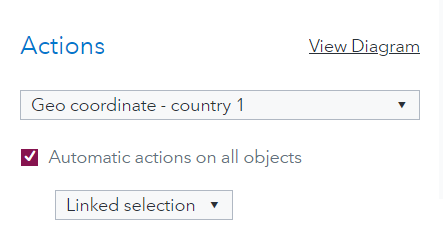
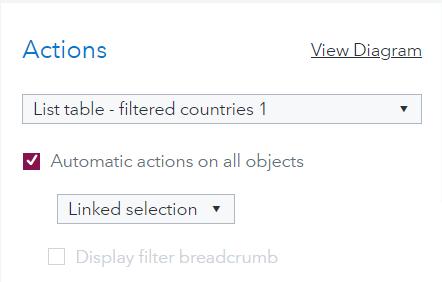
*Country, Recovered (days to double)/country, Confirmed (days to double)/country.*

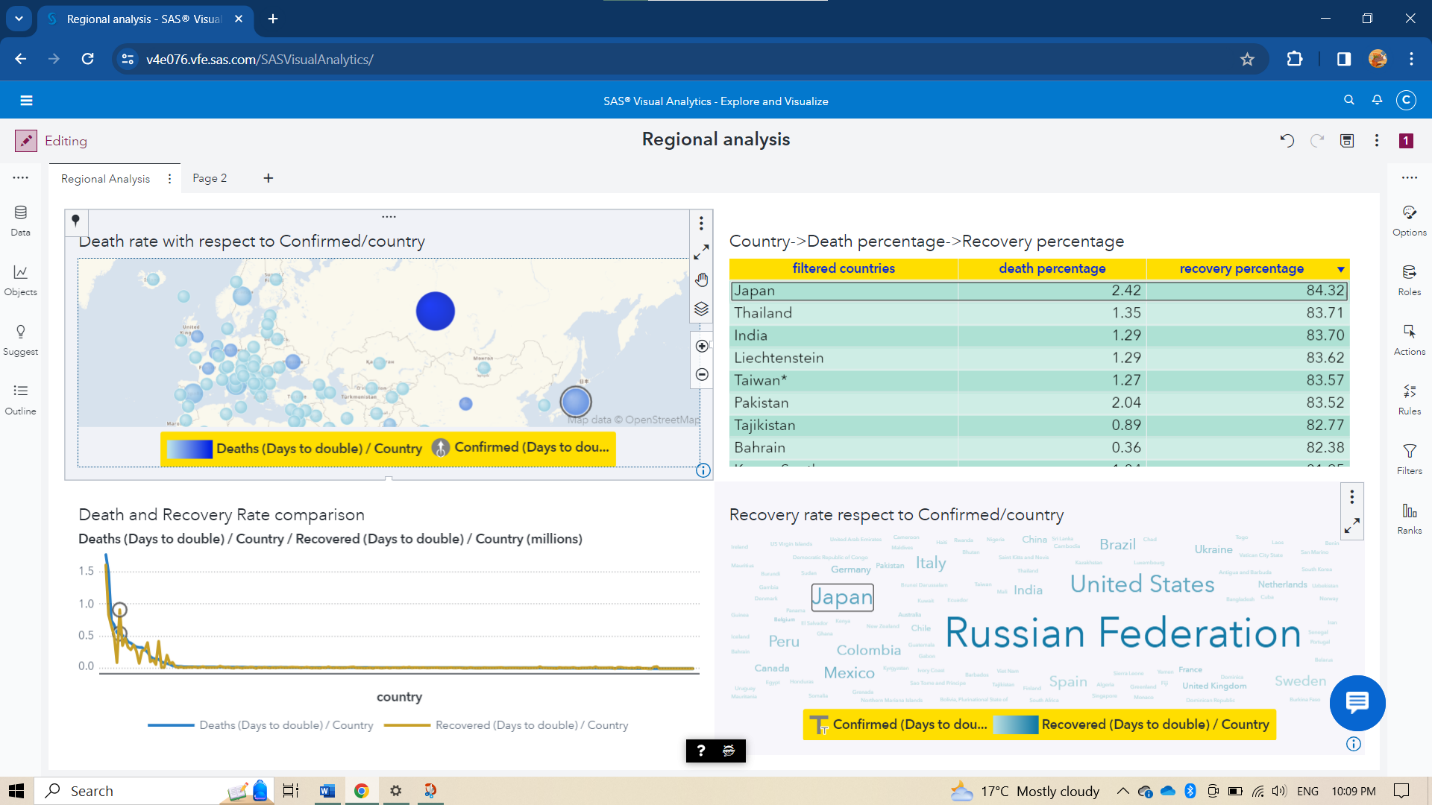
* **Insights:**

The difference between the *Recovered (days to double)/country, , Confirmed(days to double)/country* indicates the healthcare level in each country knowing that the highest difference where:

1. The recovered rate is higher -> the best healthcare level in the whole countries.
2. The death rate is higher -> the worst healthcare level in the whole countries.

****

**For the report as whole we applied Links between all the graphs**

 **THE OVERALL REPORT INSIGHTS**

1. **Japan:** is the country with the highest healthcare level concerning the NOVEL\_CORONA\_VIRUS dataset on SAS Viya platform.
2. **China:** is the country with the worst healthcare level concerning

**Certificates & Badges of course completion:**