# Covid-19 data - VisDS21

Jonas Kernebeck, Cem Kozcuer, Felix Lehner

07/16/2021

## Contents

**1 Introduction** **2**

**1.1 Covid cases and HDI by countries** **3**

**2 What are differences / commonalitites between continents?** **4**

2.1 Comparison of development indicators of selected continents . . . . . 4

2.2 Comparison of health specific indicators of selected continents . . . . . 5

**3 How do covid-19-cases and deaths differ in the countries of the EU?** **6**

3.1. Are covid deaths and cases figures different in the EU for countries with low and high GDPs in absolute and relative numbers? . . . . . 7

3.2 How are relative cases and death related to each other for EU countries with high and low GDP? 9

**4. Is there a distribution in the variables to be found in the data set?** **13**

**5 Can we detect regional differences between the continents with a PCA?** **17**

5.1 Scree Plot of PCA's for different continents . . . . . 17

5.2 Correlation BiPlot for PCA of Europe . . . . . 18

5.3 Correlation BiPlot for PCA of Europe . . . . . 19

**6 Was the data collection and compilation made meaningful?** **20**

**7 Conclusion** **26**

# 1 Introduction

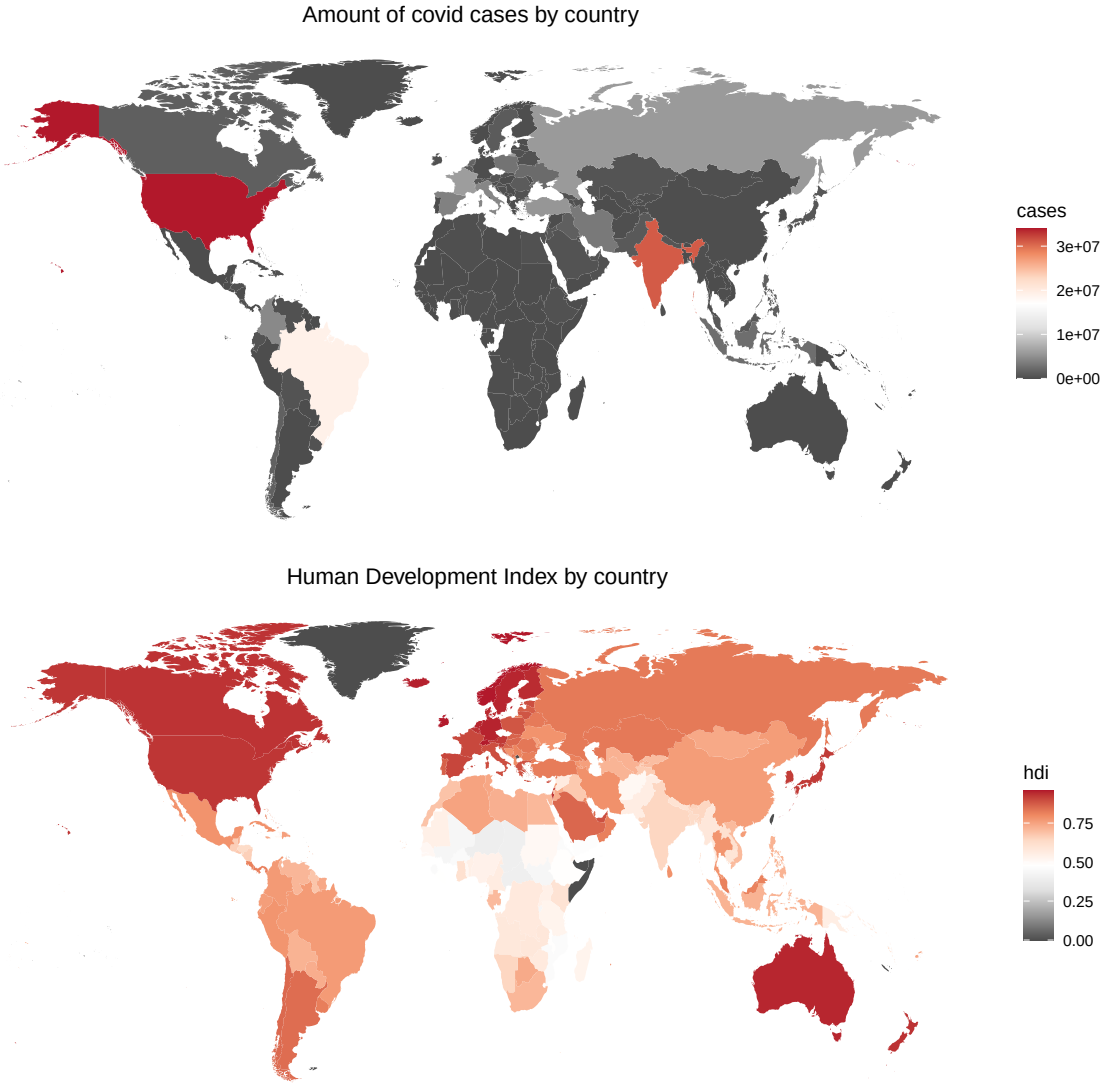
In our project we used the *COVID-19 dataset* maintained by *our world in data.*(link)

The data is mostly feed by sources from the John Hopkins University and the European Centre for Disease Prevention and Control. But also from other sources.

It is updated daily and contained roughly 100,000 samples the last time we tapped it (2021-07-11). It contains data about roughly 190 countries for everyday with 60 attributes. The attributes span different topics like data about locations, cases, deaths, tests, vaccination, hospitals, demography, medical aspects, development and economy.

In our approach we start in 2. with on overview about the different continents in regards to cases, development and health specific indicators. In the following in 3. we look closer to the EU and especially compare countries with high GDPs with low GDPs for covid-19-cases and deaths figures. In 4. analyse the distributions of the total cases and total deaths. With 5. we look at how much of the variance in the data can be explained by attributes differentiated by continents. And lastly, 6. we inspect the data set for missing entries compared on continents but also with a closer look onto Africa and Europe.

## 1.1 Covid cases and HDI by countries



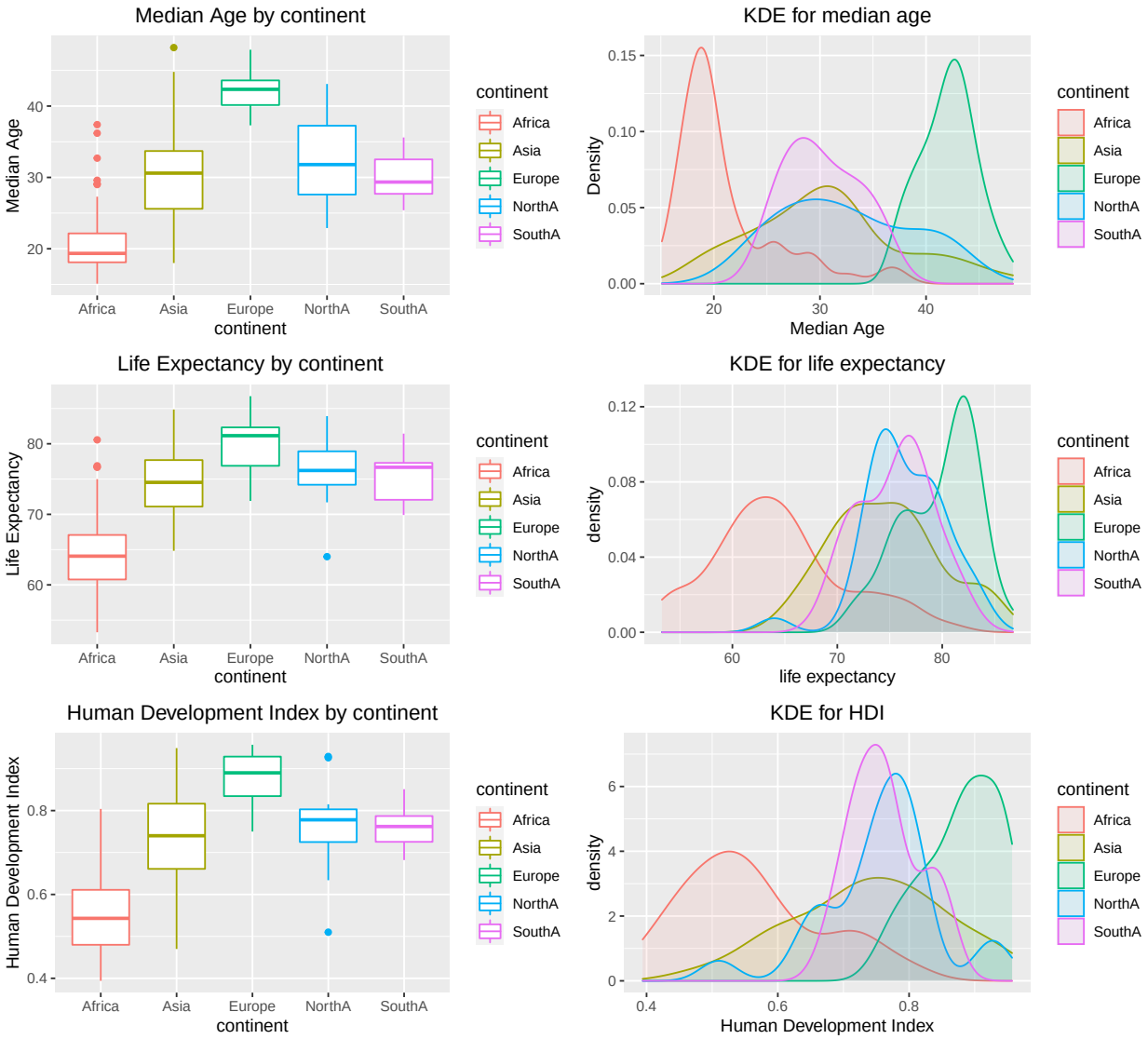

The two world maps give an overview of the amount of covid cases and the HDI by country. In absolute values, the United States and India have the most covid cases. The United States also have the highest HDI beside some countries in central / northern Europe and Australia. This shows, that the dimensions of the Covid-19 spread is not only dependent from the development of a country, but also from other important factors. Therefore, we want to focus on the differences and commons of some development and health indicators between continents. After that, a country-specific comparision will be made.

# 2 What are differences / commonalitites between continents?

For the continent-specific comparison of different development and health indicators, we look at boxplots and kernel density estimations for different variables. All of these variables are time-independent and are the values for the year 2020 of the countries respectively. The different boxplots display the ranges for the different continents.

## 2.1 Comparison of development indicators of selected continents

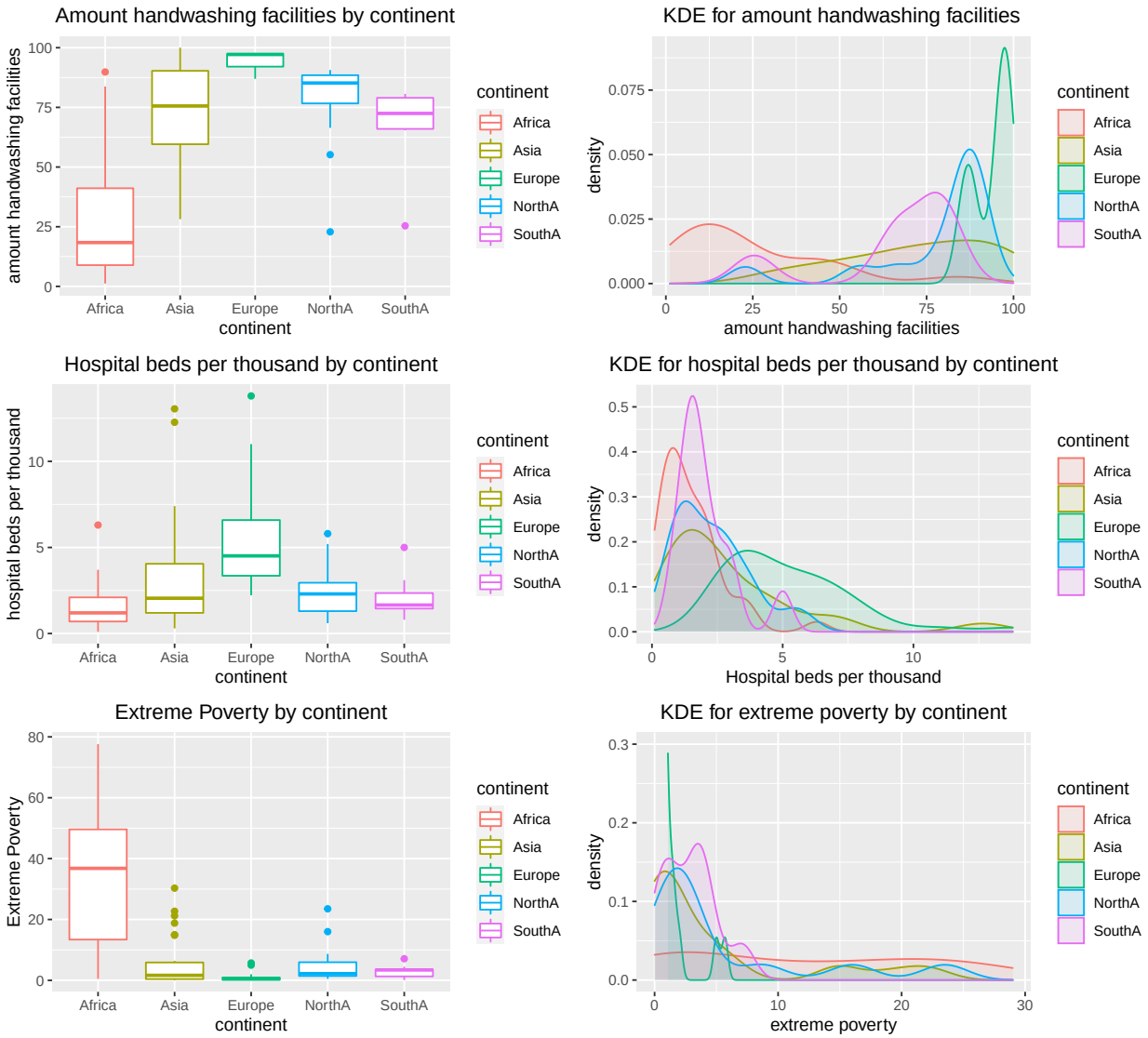
For the development indicators, we look at the median age, the life expectancy and the already considered human development index.




Conclusion: For all of the three indicators, Europe has always the highest level. The average and the third quantile is higher than all of the other countries. It is also noticeable that the range between the countries in Europe is always small compared to continents like Africa and Asia. Africa has the lowest niveau of all three development indicators. It also has some outliers in the median age and the life expectancy.

## 2.2 Comparison of health specific indicators of selected continents

For the health-specific indicators, we look at the amount of handwashing facilities, amount of hospital beds per thousand and the extreme poverty.

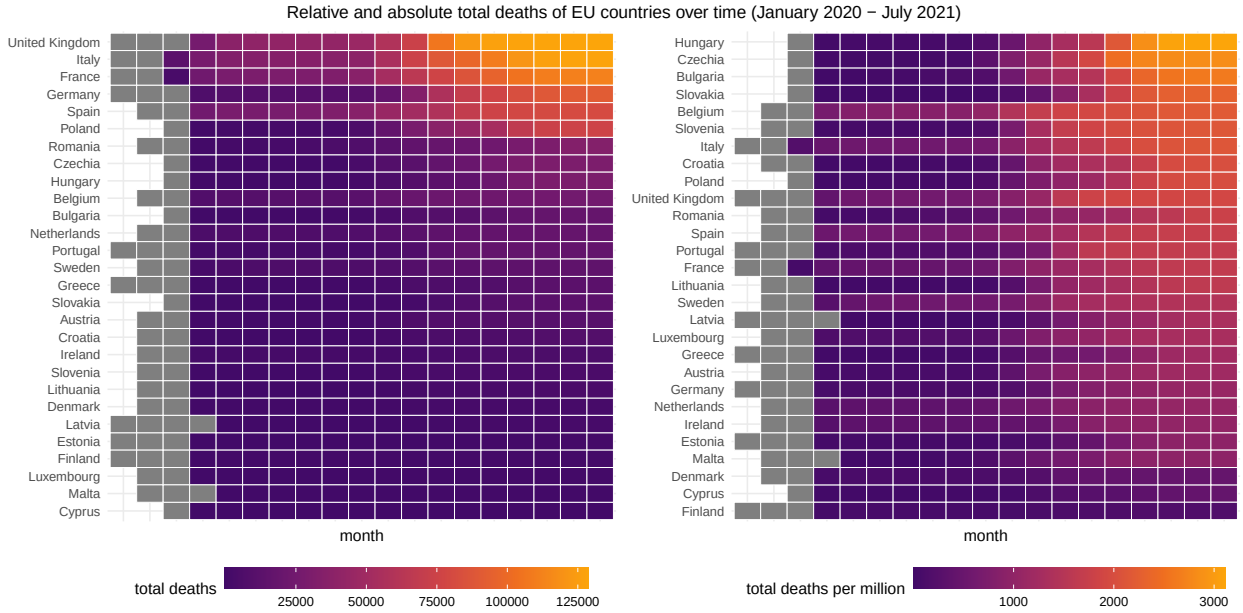



Conclusion: For the health-specific indicators, we can detect a similar tendency as in the development sector. Europe has similar good results in the health sector with the lowest amount of countries with extreme poverty and the highest amount of handwashing facilities and hospital beds per thousand. Also the range within the continent of Europe is the lowest for the handwashing facilities and the extreme poverty. The continent Africa has the highest range in handwashing facilities and extreme poverty and also the weakest level of all continents.

# 3 How do covid-19-cases and deaths differ in the countries of the EU?

In order to get an overview over the covid-19-cases and covid-death figures in the EU we first listed all countries in heatmaps regarding *total deaths*, *total cases*, *total deaths per million* and *total cases per million* figures. Both absolute and relative figures are shown over the time of the pandemic and sorted by the maximum of the attributes.

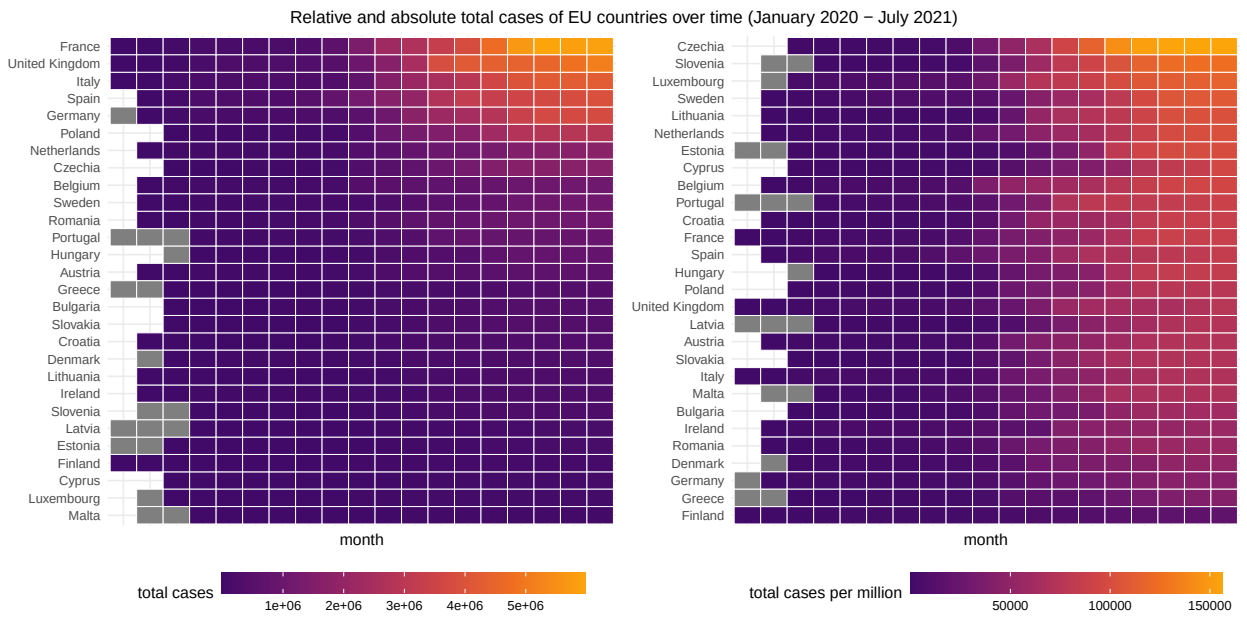
The heatmap can show nicely how the total figures for each country develop over time for each country. Without any overlappings the countries are easily to compare with each other. The color coding which shows the accumulation of the numbers indicate how high the numbers of each country are and how fast they raised.




Conclusions: Compared to the absolute death figures the relative figures show a different picture. The countries with the most total deaths are differently from each other when looked at the relative numbers:

Whereas Germany is rather on the lower third regarding relative deaths, United Kingdom, Italy and Poland are on the topper third. Other countries as France and Spain are to be found in the midfield regarding the relative numbers. On the other hand side countries with lower absolute deaths have high relative numbers.

Further in the heatmap for the absolute deaths there is also a cut to be seen after the top 6 countries: We would assume that that has to do with the population jump of the larger EU countries to the other countries. We can also denote that the distribution of the relative numbers over time are much more even than on the total numbers, which also affirms the assumption that absolute deaths numbers are tight to the population amount.



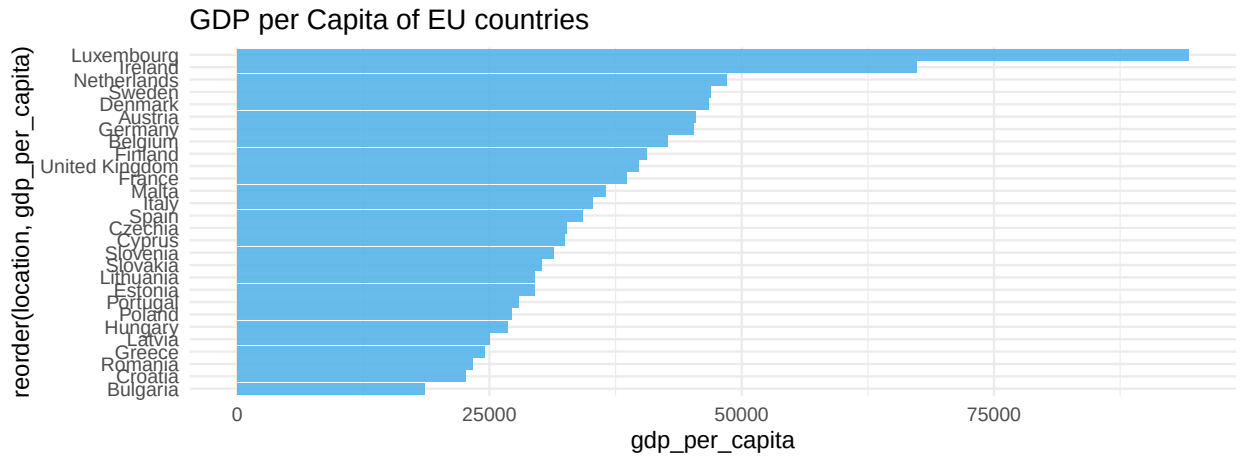

Conclusions: Looking at the covid-19-cases shows an equally picture for the absolute figures as like deaths. But interestingly not only the relative numbers are different for a country compared to the absolute.

There is also a difference between relative deaths and relative cases:

Countries like Sweden, Lithuania, Netherlands and Luxembourg have a very high relative case number but low relative death number. In this countries the cases lead to much less deaths than in countries like Hungary, Czechia or Bulgary.

## 3.1. Are covid deaths and cases figures different in the EU for countries with low and high GDPs in absolute and relative numbers?

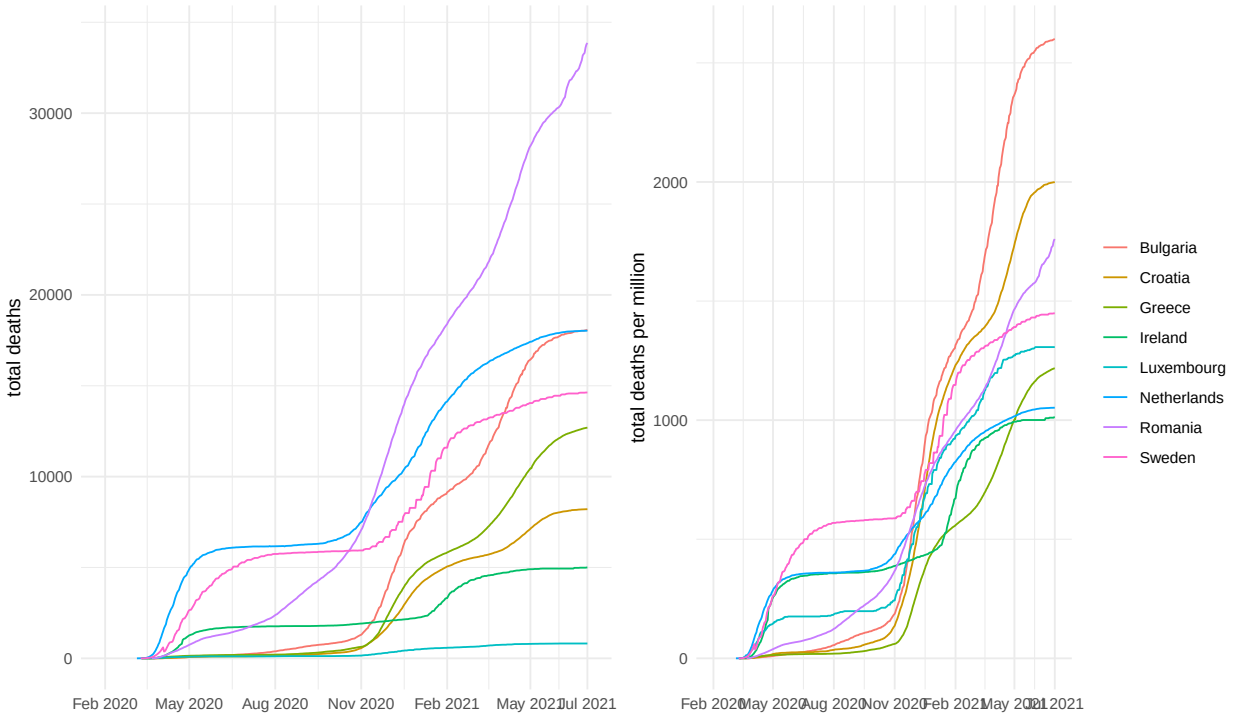
The heatmap overview gives indications about the "richness" of the EU countries being involved into covid cases and deaths numbers. So we looked at the *GDP per capita* values for each country in order to select countries with high and low GDPs and to look at the differences.



**Absolute and relative total deaths of EU countries with low and high GDP over time**

With the line plots here we compared absolute and relative figures of *total deaths* and *total cases*.

The total numbers over time are technically a cdf of the deaths and cases. So with the line plot we can see for each country not only how high the values are compared to the absolute values but also the comparison to the other countries. The gradient of the curves shows how fast for each country at which time point the covid-19-cases and deaths develop. Basically steep parts of the curves show raising infections and deaths and plateaus stagnation of the numbers.

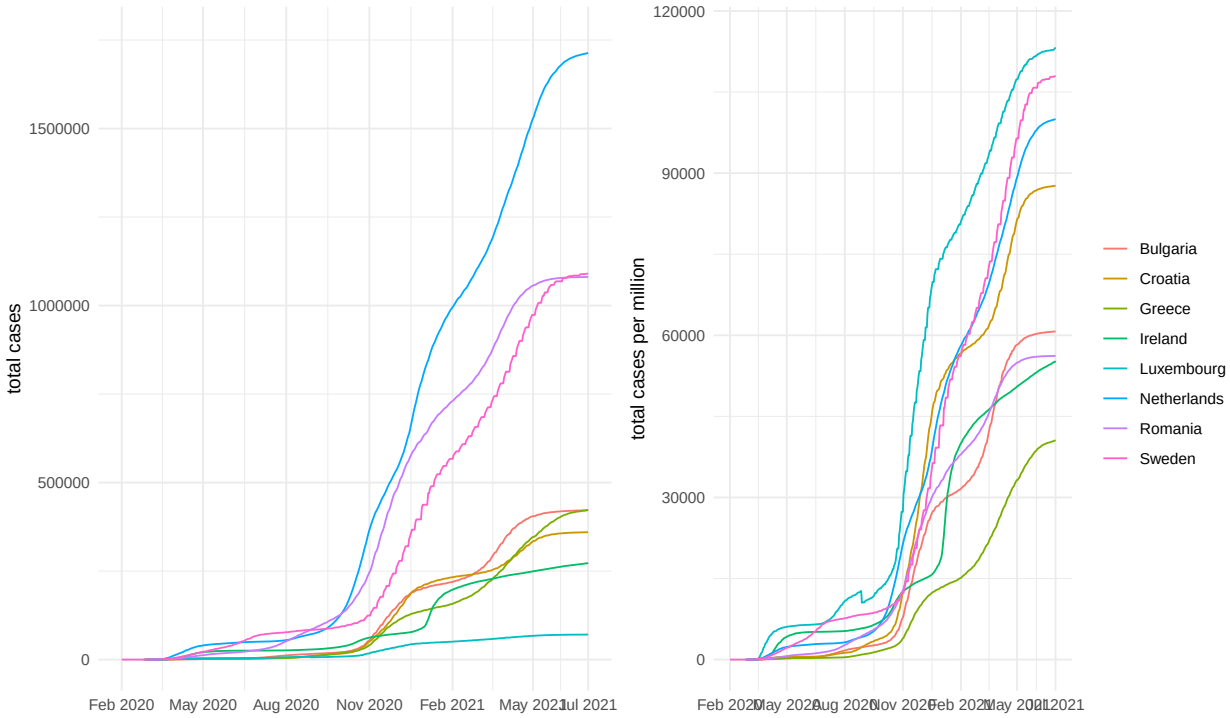



Conclusions: As the heatmaps also showed here the picture is subtly differentiated: Luxembourg (also Ireland) and Romania (also Bulgaria) are the extremes: Luxembourg has both low death in absolute and relative figures. Whereas in Romania both are high. Sweden and Greece are both quite in the midfield, even they are opposite in their GDPs.

In comparison to the heatmaps, we can see here that there were some time points where the figures raise very quickly: At the beginning of the pandemic and in November 2020. At the moment it seems we are getting again to a plateau, which might be related to vaccination.

Romania however has a constant raise in deaths both in absolute and relative numbers, which implies that there is effectively no measures against covid-19-deaths.

**Absolute and relative total cases of EU countries with low and high GDP over time**




The cases plots look a little different. The absolute and the relative numbers seem to more correlate positively with each other. Like on the heatmaps we can see that Luxembourg is quite extreme: low overall cases but high relative numbers. But compared to the relative death figures: lowest mortality. Bulgaria for example show in that regards a much higher mortality by cases. Also again we can see very extreme raise of figures from November 2020. And a plateau currently.

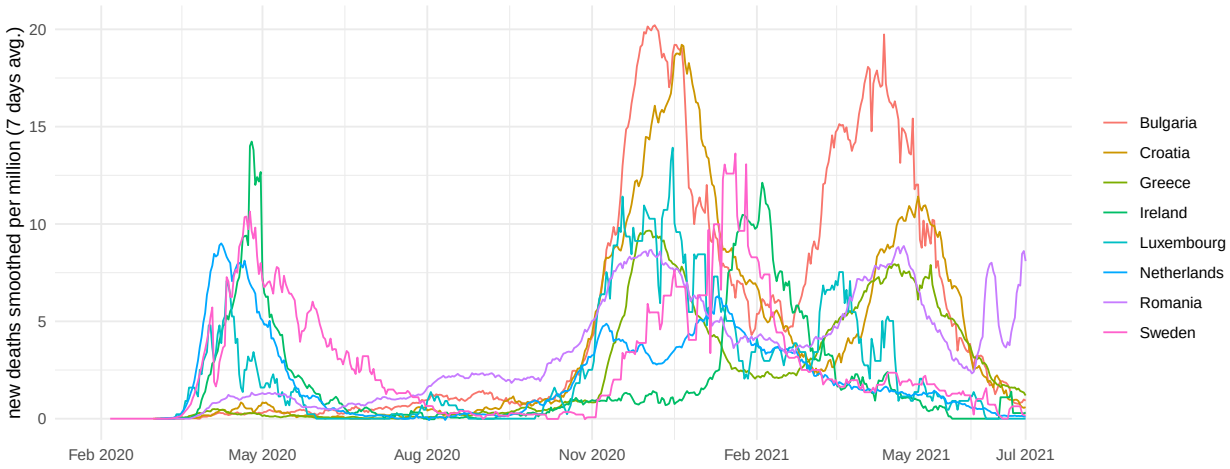
November might be related the beta variant of corona. And the current plateau to the effect of vaccination.

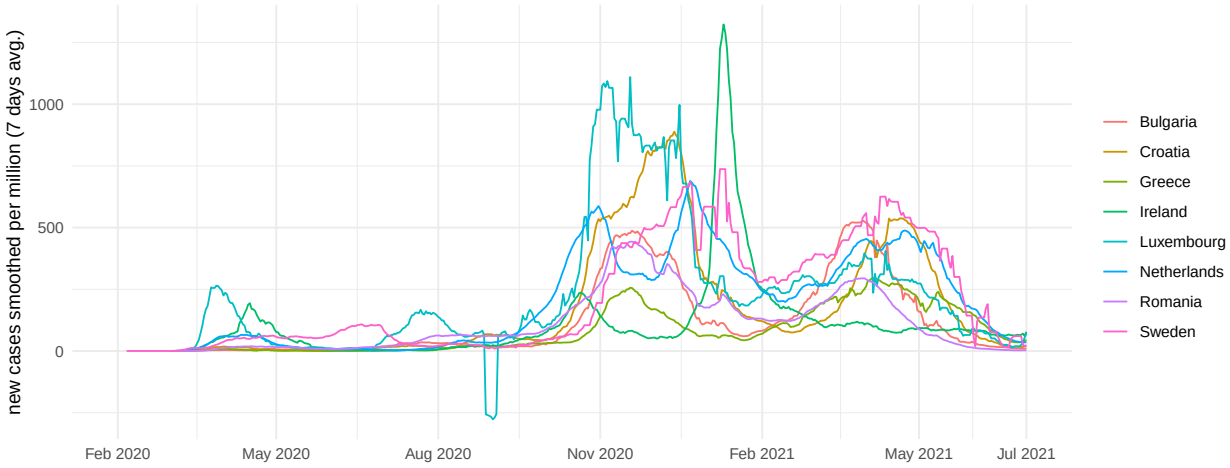
For Romania compared to the death figures we can here see also the figures developing towards a plateau, which might also have an effect on the constant deaths in the close future.

## 3.2 How are relative cases and death related to each other for EU countries with high and low GDP?

We have seen so far relations between covid-19-cases and deaths but also that they are subtly differentiated from country to country.

In order to get more insight about the relation of cases and deaths we looked at the relative *new cases* and *new deaths* compared between the countries.




Conclusions: The picture is again different from country to country. There is definitely a correlation between cases and death to be seen. But it seems not be dependent on the GDP.

Compared to the heatmap we can see here on the time axis the corona waves. Each wave causes deaths. But a high GDP country like Sweden has quite a lot deaths for relatively low cases in the beginning of the pandemic but very few death for quite a lot infections in spring 2021, which might be related to different reactions, countermeasure or vaccination.

But this picture is different for Bulgaria: Almost untouched from the pandemic in the beginning Bulgaria has very high relative deaths at the second and third wave similarly like Croatia, maybe due to a later availability of vaccination.

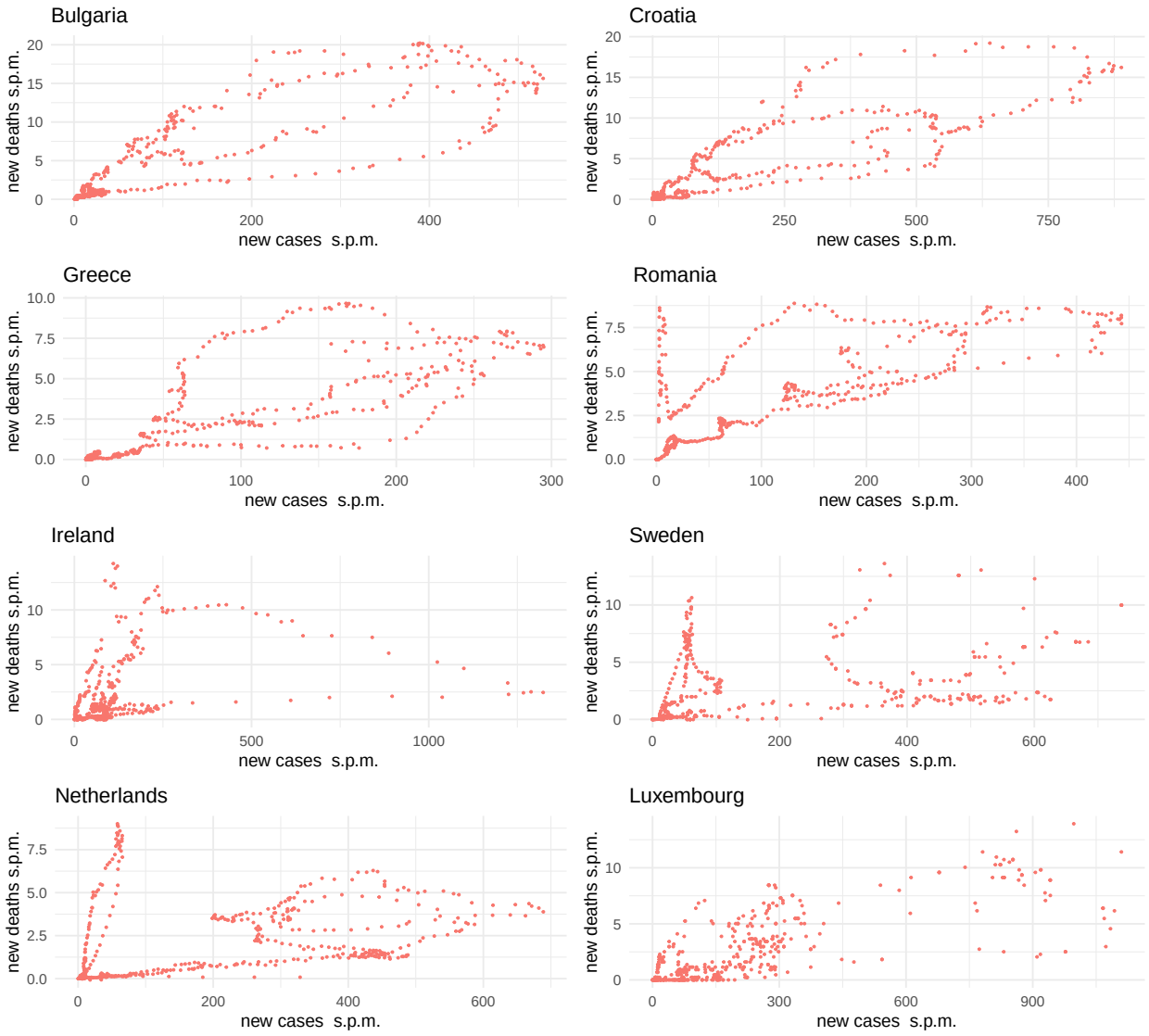
The last wave is almost not existent for high GDP countries whereas for the low GDP countries it is quite visible. Except of Romania these countries however seems to managed the third corona wave and have currently similar new deaths figures as the high GDP countries. This might be related to later availibility of vaccination but of course also to other measures in the countries as contact and mask regulations.

For the new cases rate we noticed also that Ireland had very low infections through out the third wave. Romania however with relatively low case rates has quite high new deaths numbers.

**Correlation between relative deaths and relative cases per country**

In the following to examine the correlation between *new deaths* and *new cases* we used a scatter plot where the two variables are plotted against each other and the points are grouped by countries.


Conclusion: The plot looks a bit wild, but it is visible that the different countries differ from each other: Some seems to have lower relative deaths on relative cases than others. So a linear relationship per country is indicated with different steepness of the directions of the point clouds per country. In order to understand the differences better between the countries we did split the plot by countries so we could analyse the shape of each countries point cloud in comparison.



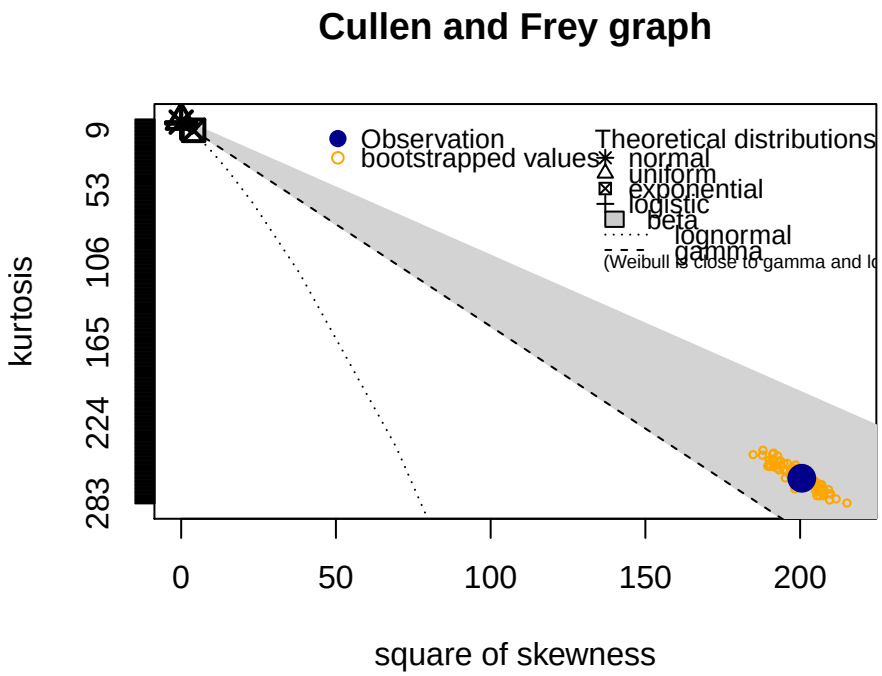

Conclusion: The point clouds are not easy to interpret, but still reveal some interesting patterns: Netherlands, Luxembourg, Sweden, Ireland (high GDP) have quite flat curves - but all some big comparable spike, which low GDP countries seem not to have. Maybe the beginning of pandemic had more impact to high GDP countries. (Because of more transits?). For the low GDP countries there are 3 "different paths" recognizable in the plots. These might correspond to the 3 big waves that maybe had different intensity in new death figures.

## 4. Is there a distribution in the variables to be found in the data set?

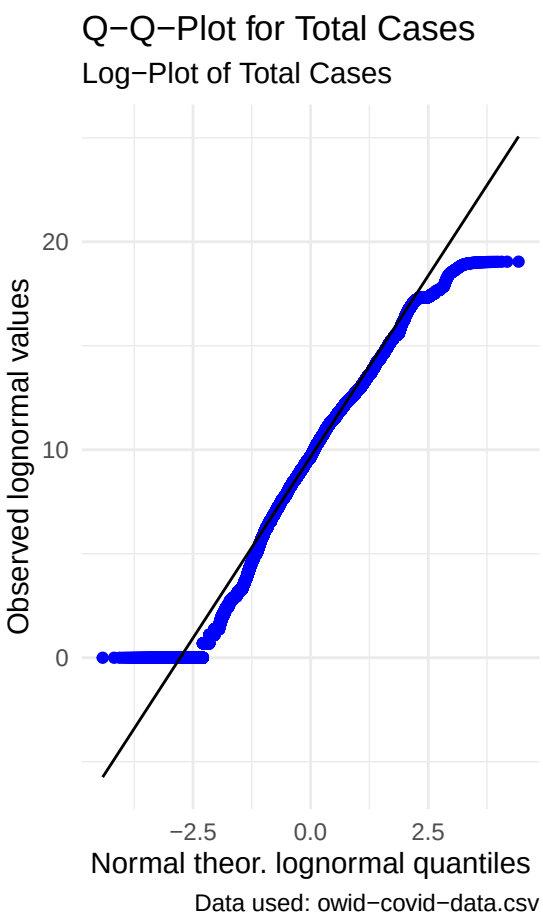
In order to analyze whether there is a target distribution within the variables, the most common and best known variables - Total Cases, Total Deaths and Reproduction Rate - are selected to be investigated. It should be mentioned that the data are time-based and that time series analyzes should normally be used. However, we have omitted this chapter for reasons of content.

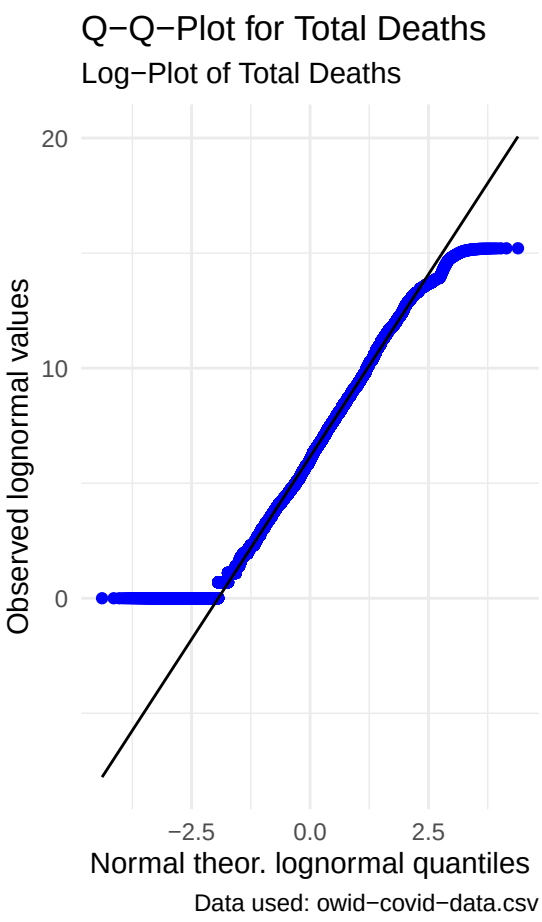
In the course of the investigations the function descdist() from the package fitdistrplus serves mainly. This compares several Pearson distributions with the variables we know and additionally introduces bootstrapping.

**Total Cases and Deaths**




Conclusion: This plot shows only the disributions for total cases, as for total deaths it is almost the same. The variables total_cases (total corona cases) and total_deaths (total corona deaths) can be fitted to a lognormal distribution. It can also be fitted to a gamma function, but since it is known from observation that these variables are cumulative and increasing, only the lognormal distribution is examined.



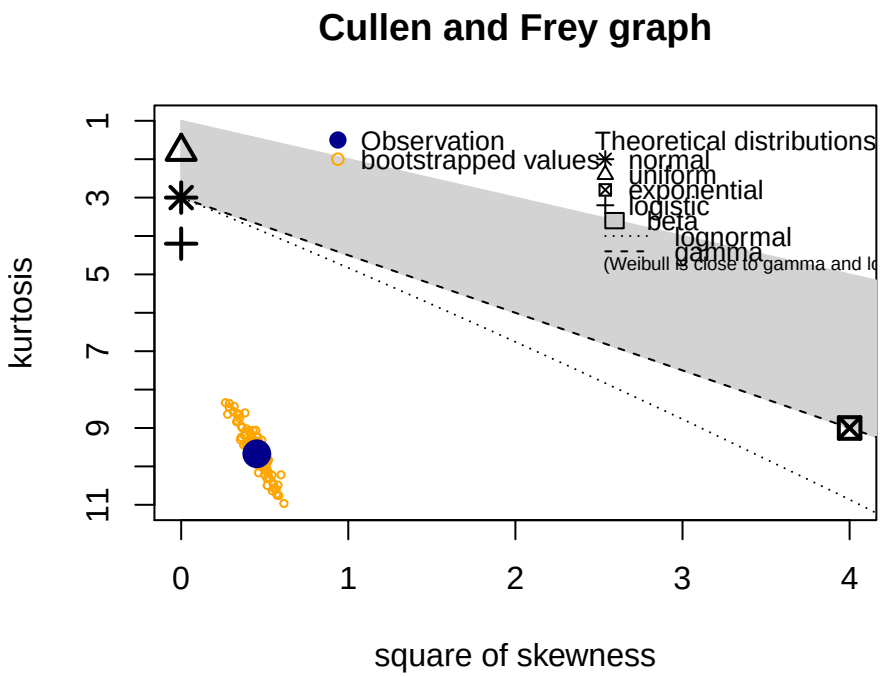




Conclusion:

- Total Cases: Except of the tails of the plot, the data looks very lognormal distributed. The tail at the bottom is understandable, because there are no negative cases. The little outbreak at the top of the graph indicates that the observations are increasing that much like a lognormal distribution would expect.
- Total Deaths: The tails are also not lognormal like in the total cases analysis, but the general curve is closer to the theoretical distribution. All in all the distribution is heavy tailed and have the same characteristics as the total cases variable.

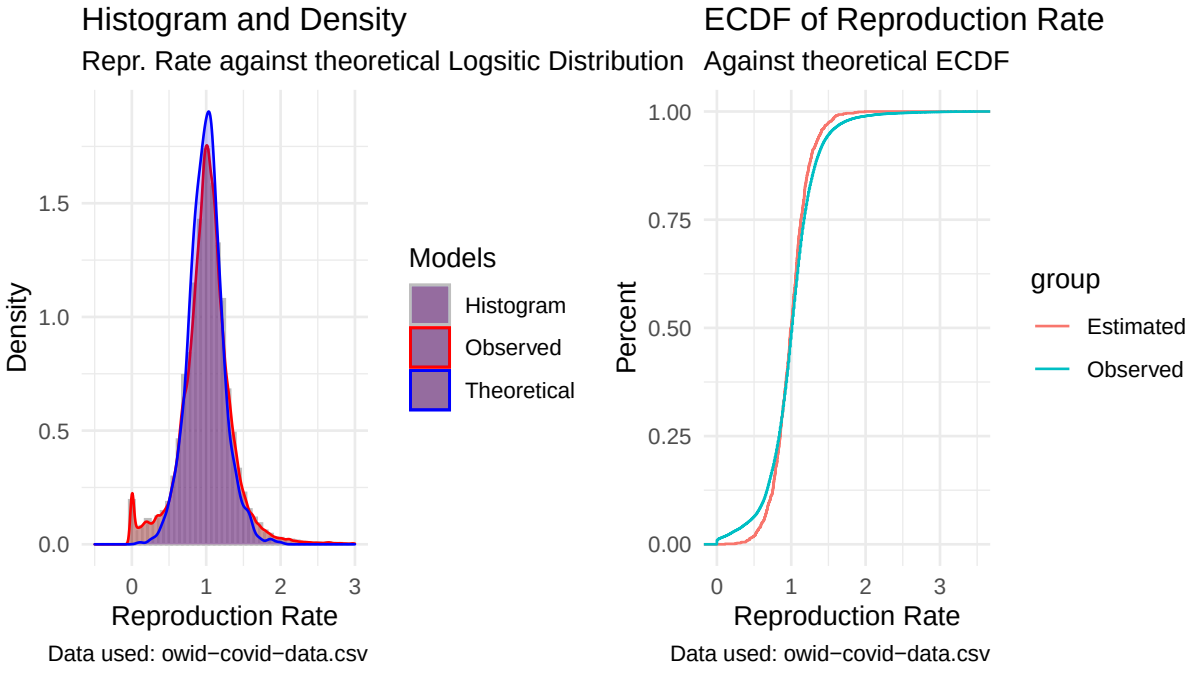
**Reproduction Rate**

As expected the total cases and deaths are lognormal distributed, which is understandable, because it is a value which is always increasing. Another more interesting variable is the reproduction rate.




Conclusion: At first glance, it looks as if this variable does not match any of the distribution. On closer inspection, however, it becomes clear that the reproductive rate exhibits a logistic distribution approach.

In order to analyze the distribution more precisely, the density is compared with the density of a theoretical logistic distribution and an empirical density distribution function.



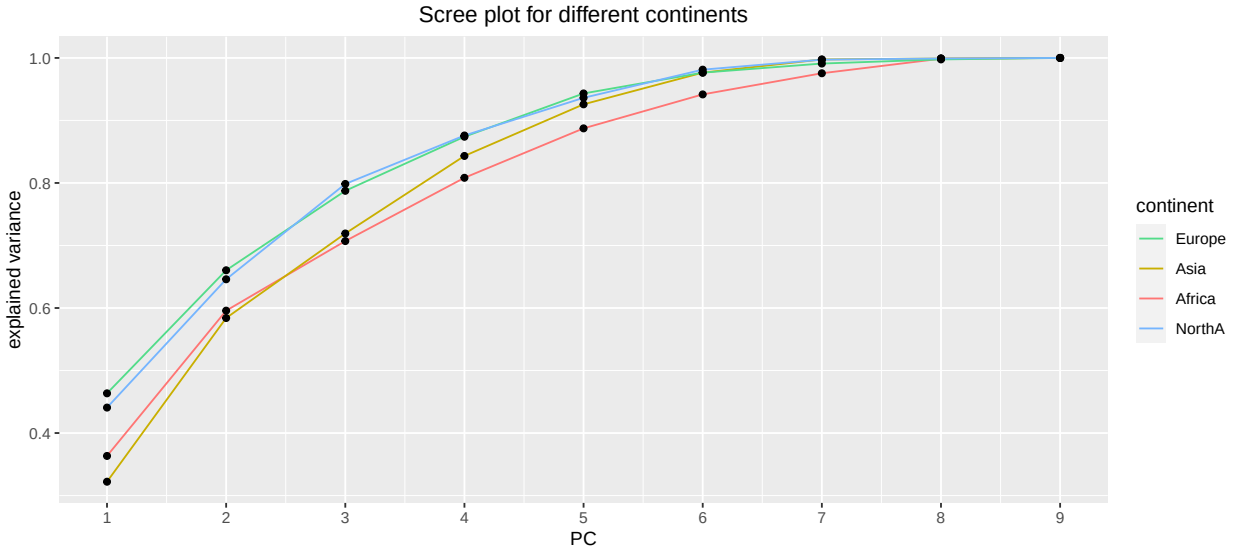
Conclusion:

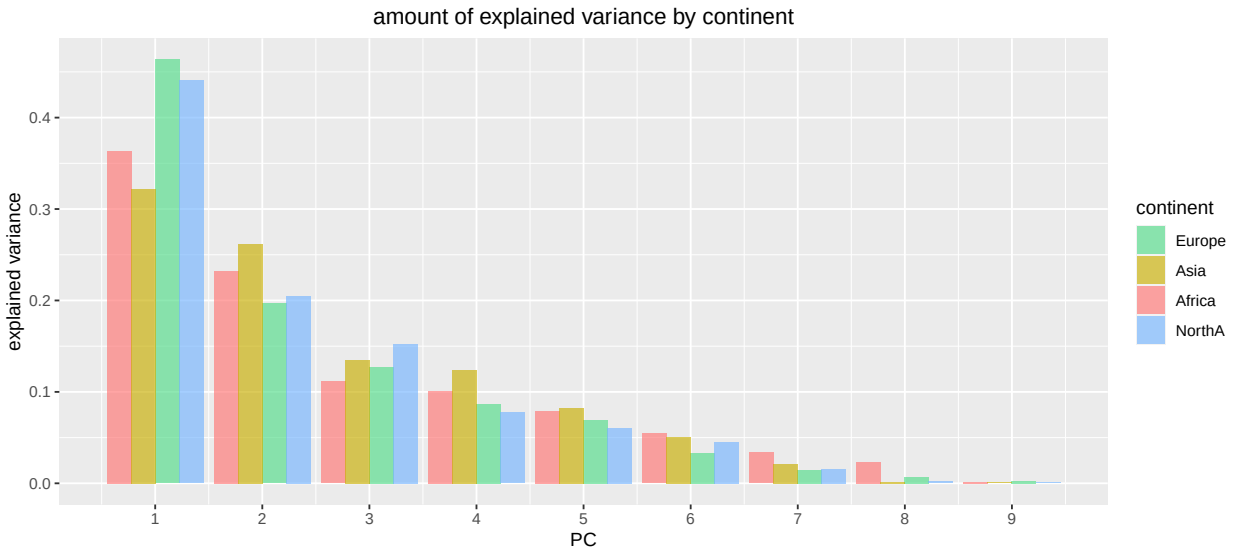
- Density: The observed data of the reproduction rate is fairly logistical distributed. Some outliers, especially around zero, disturb the distribution. The observed data has wider tails and a lower tip than the theoretical one.
- ECDF: The observed data of the reproduction rate is also here nearly logistic distributed. The theoretical data is also steeper than the observed one.

# 5 Can we detect regional differences between the continents with a PCA?

## 5.1 Scree Plot of PCA's for different continents

To detect regional differences between the continents, for the continents Africa, Asia, Europe and North America a PCA was made. The following Scree and Boxplot show the different amount of explained variance by the principal components.




Conclusion: The results of the different PCAs show differences between the continents. The curve of the the Scree Plot for the continents Africa and Asia run similar. The same effect can be seen for the continents Europe and North America. Beside the similar curve of the two continent-pairs, there are also differences between the amount of explained variance per principal component. While the first component for North America and Europe explains about 44% and 46% of the total data variance, the component for Africa and Asica explains only about 36% and 33%.

## 5.2 Correlation BiPlot for PCA of Europe

To find out, how the different principal components of the PCA of Europe are build, we look at a correlation biplot. The plot shows the share of the variance of the different variables on the different principal components.


Conclusion: We can see, that the first dimension of the PCA contains nearly only covid data and the population. Development data plays no crucial rule here. The whole variance of the development data is mostly distributed on the principal component 2 to 5. With the first two principal components for the PCA in Europe, nearly 70% of the total variance can be explained. This shows, that the whole data variance can be approximated with covid and population data, because these explain the differences between countries in Europe the best.

## 5.3 Correlation BiPlot for PCA of Europe

To find out, how the different principal components of the PCA of Africa are build, we look at a correlation biplot. The plot shows the share of the variance of the different variables on the different principal components.

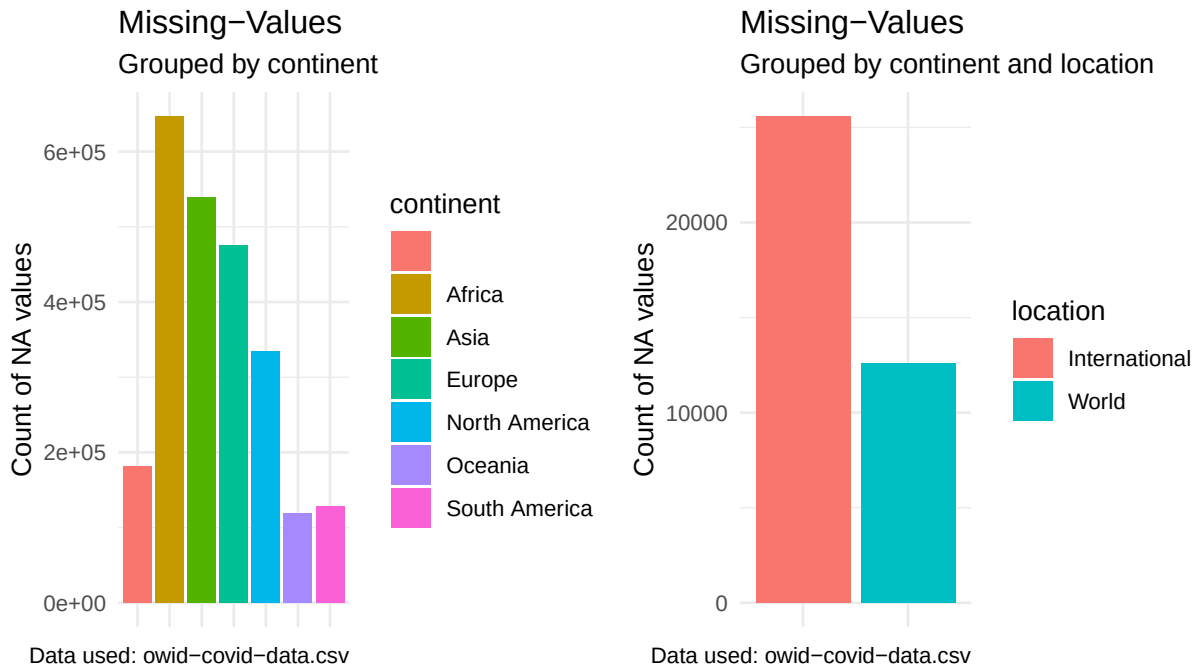

Conclusion: In contrast to Europe, the first dimensions of the PCA of Africa contain the great part of the total variance of the development data. The covid data plays , like in Europe, a crucial role. The differences between countries in Africa can be explained with covid data. Beside this data, also the development indicators explain a great part of the total variance, because of the disparities between the countries.

# 6 Was the data collection and compilation made meaningful?

To examine the NA values in the data set, they are added together in rows. This puts the consideration on the observations and not on the missing values per column, since these are already to be found in the summary of the data set. Accordingly, another column is created in which the number of NA values per row are stored.

### Grouping by Continent

To get a first overview of the missing values, they are illustrated by a barplot, which is grouped by continents.

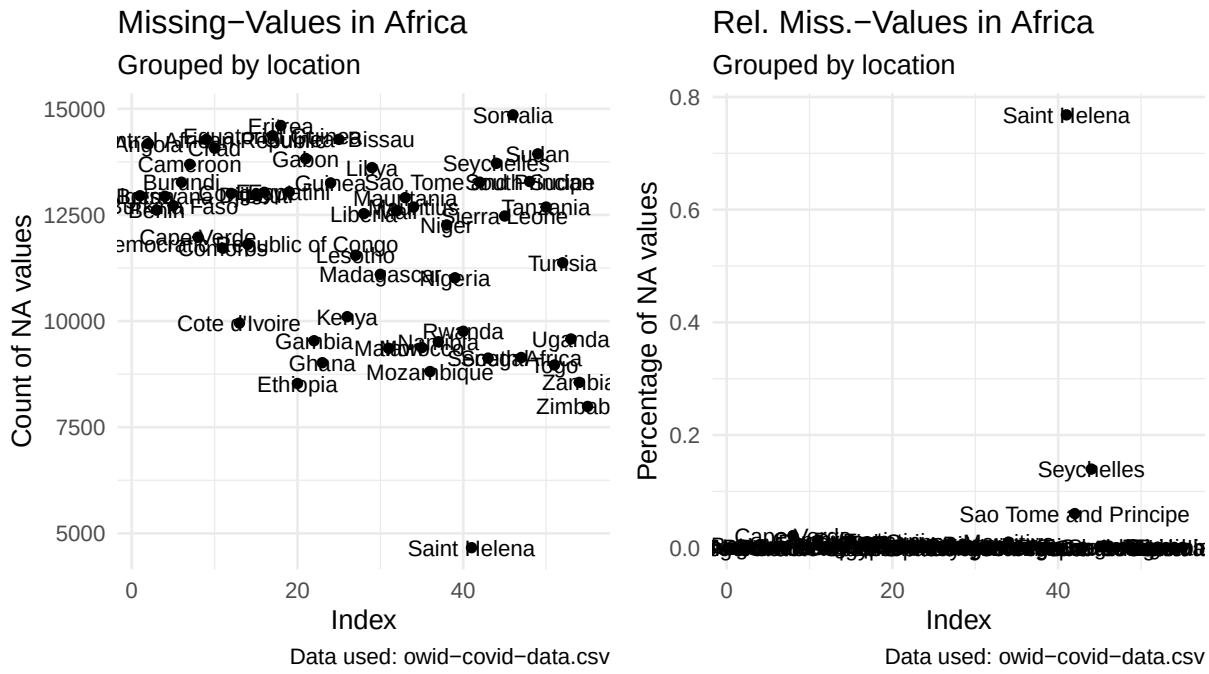



Conclusion: In Africa, Asia and Europe the most NA values are to be found. There is one continent that has NA values but finds no affiliation. To get a deeper insight into the 'Non'-continent, the right figure shows the dataset which is additionally grouped by location and the other known continents removed. International and World are also present in the dataset and presumably show a general mapping in the dataset. In the next analyses, International and World will be separated out and not considered further.

### Compare Africa and Europe (ScatterPlot)

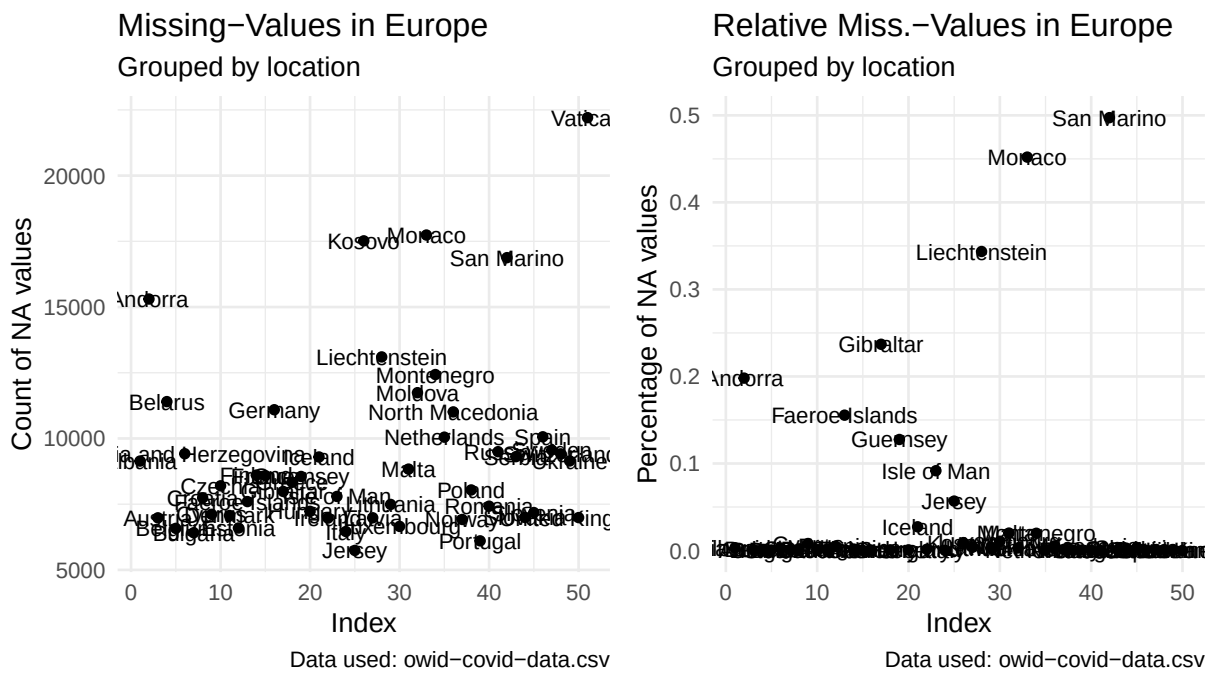
Based on the results from the previous chapters, we will now look at the two continents of Africa and Europe. The individual locations in each country will be compared.

The first continent to be considered is Africa and its places. For this, a scatter plot is selected to see which country has the most of the missing values.




Conclusion: In the left view Saint Helena has the fewest missing values in the data, but it is also one of the smallest countries in Africa. The data needs to be scaled to see what country has the fewest or most missing values per population. The right view shows the scaled data. Saint Helena is now the country with the most missing values. This indicates that although Saint Helena has less missing values the missing values are much more severe than in the other countries of Africa.

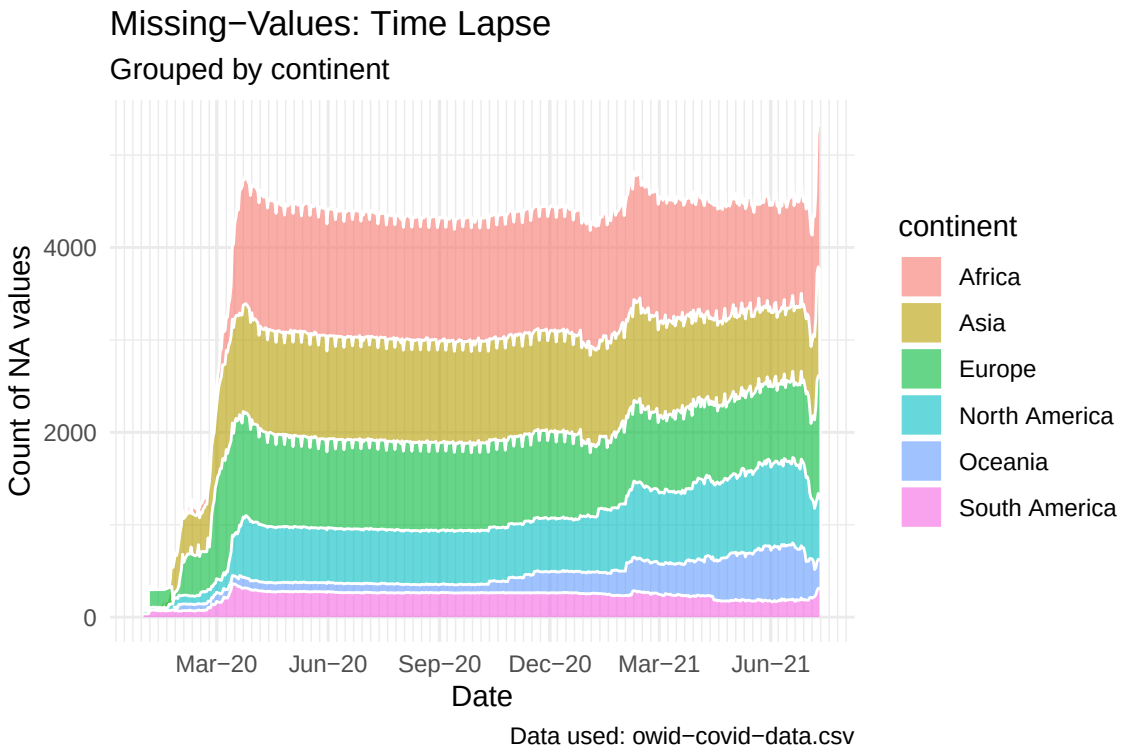
The next continent to be analyzed is Europe. For this the same procedure is done.




Conclusion: Some countries do have a lot of missing data, but in contrast to the population it is quite small. It is noticeable that some of the countries are on the same level in terms of absolute and relative values, whereas others a slided to the bottom and some to the top of the relative view.

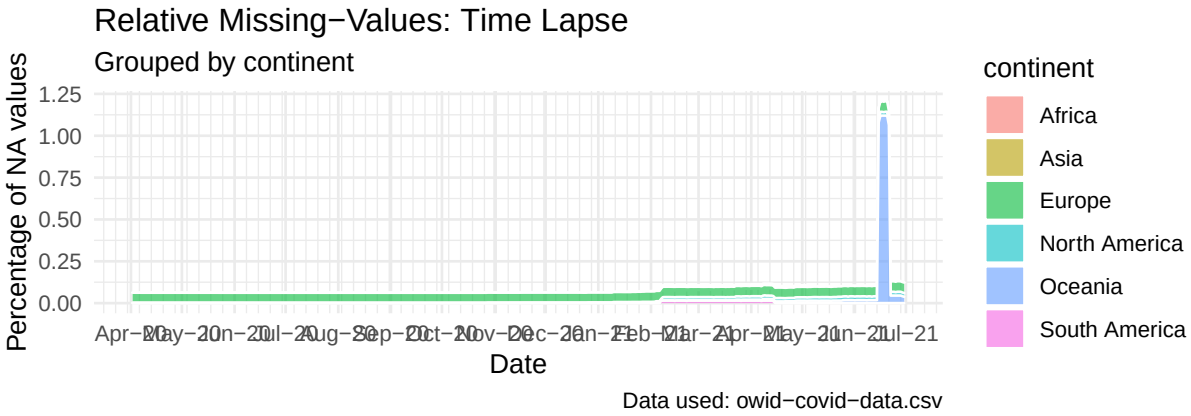
**Time-Lapse: NA-Values per continent**

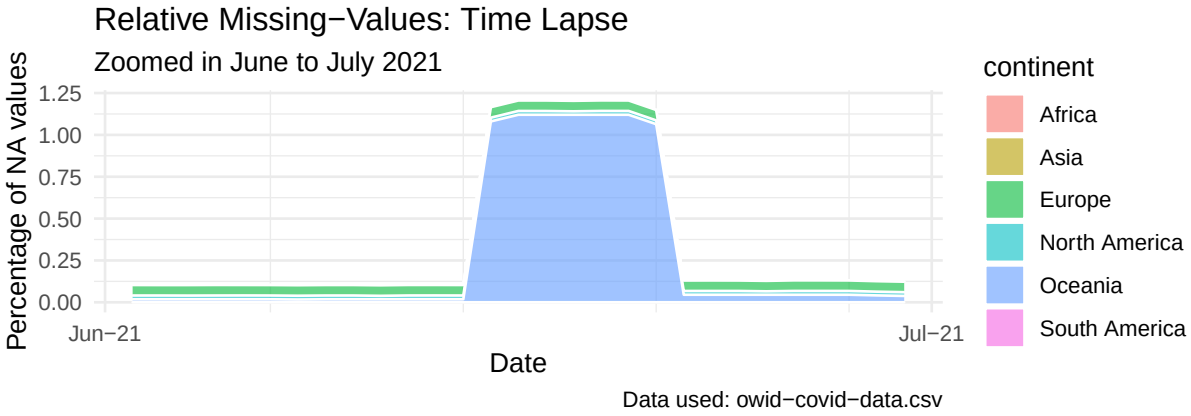
The next step in the analysis is to visualize the missing data over time. Start is January 2020 and end is July 2021. For this a area plot is used to identify outliers and changes in the continent view.




Conclusion: There are some special outbreaks that are amplified per connected continent. In general the missing values remain constant in the data collecting process though the time. Oceania is a continent in which the missing values increase from December 2020 onwards.

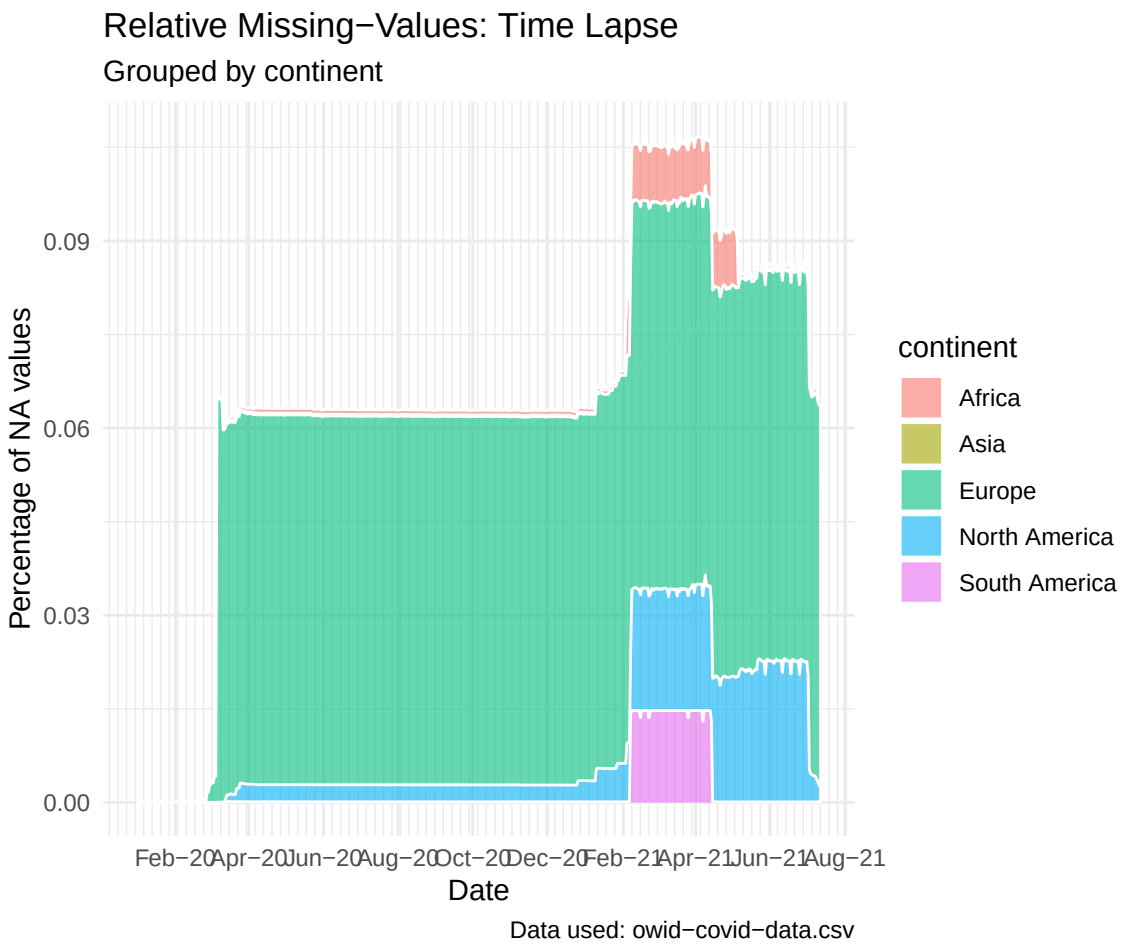
The same process can also be done with the Ratio of NA_values per population, which was an introduced parameter from the previous chapter.




Conclusion: The outbreak from Oceania between June and July 2021 destroys the plot of the Time Lapse (there is something going wrong).

To clean the data and the plot from the disturbing Oceania data, the continent will simply be filtered out of it.




Conclusion: South America does have an interesting leak of data from February 2021 to April 2021, which can be investigated more.

# 7 Conclusion

Overall the covid-19 data set is a very rich, but also challenging data set. Our data exploration gave us a lot of indications of how the covid-19-situation has a different impact on different continents and countries and how it leads to different figures regarding the mortality.

In 2., the global overview, and in 5., the PCA-analysis, we showed that different attributes seem to be differently relevant to different locations.

Looking closer onto the relation of death and cases in 3., and 4. we could derive that economy seems partly related to the death and case rates, but could by far not explain everything. Over time the pandemic has a different impact on the countries and also seemed to change their reactions and strategies regarding new covid-19-cases. Our conclusion here is that every country is very special regarding its dealing with the pandemic and needs tailored solutions.

We are also aware that the pandemic is a very dynamic situation where the regulations regarding contacts and hygiene measure were very different for different countries and even changed over time. Tests and vaccination where differently distributed and used in the countries which we did not take into account. Some aspects are simply also not existent in the data.

Except that, the data collection process has also its own challenges which we showed in 6., when we looked how missing entries in the data are distributed between different continents and states.

Finally we understand the outcome of our exploration as a chance for hints for further investigation and as an assistance in focusing on different aspects for different continents and countries when looking at the pandemic.