Analysis of Covid19 Dataset

Author: Mehrdad Baradaran

Abstract

This study explores the temporal dynamics of COVID-19 related metrics across countries, focusing on excess mortality, testing rates, and socio-economic factors. Leveraging a comprehensive dataset, we employ visualizations and statistical analyses to uncover patterns and correlations, shedding light on the multifaceted impact of the pandemic. Our findings reveal nuanced trends in excess mortality and its correlation with testing rates, offering insights into the efficacy of public health measures. Additionally, we employ clustering techniques to discern distinct groups of countries based on human development indices and reproduction rates. The study culminates in a dynamic visual representation, capturing the evolving scatterplots of countries over time.

1. Introduction

The "Our World in Data - COVID-19" dataset, available on Kaggle, serves as a comprehensive resource for understanding the global impact of the COVID-19 pandemic. Compiled by Our World in Data, this dataset provides a rich collection of information that spans various dimensions, enabling researchers, analysts, and the public to explore and analyze the evolution of the pandemic across countries and regions.

Key Features:

- Geographic Coverage: The dataset encompasses a broad range of countries and regions, offering a global perspective on the spread and impact of COVID-19.

- Temporal Evolution: With a temporal scope, the dataset includes time-series data, allowing users to trace the progression of the pandemic from its early stages to the present day.

- Metrics and Indicators: It covers a diverse set of metrics, including confirmed cases, deaths, testing rates, vaccination statistics, and more. This granularity enables a thorough examination of the various facets of the pandemic response.

- Demographic Information: The dataset may include demographic details, such as population size, which can be crucial for contextualizing and interpreting the COVID-19 metrics.

- Data Consistency: Our World in Data is renowned for its commitment to data accuracy and transparency. The dataset is regularly updated, ensuring that users have access to the latest and most reliable information.

Potential Applications:

- Epidemiological Research: Researchers can leverage the dataset to conduct epidemiological studies, analyze trends, and gain insights into the dynamics of the virus.
- Public Health Policy: Policymakers can utilize the dataset to inform and assess public health policies, vaccination campaigns, and mitigation strategies.
- Data-driven Decision Making: Analysts and decision-makers can employ the dataset to make informed decisions based on a comprehensive understanding of the global and regional COVID-19 landscape.

2. Data Loading & Overview

Let's first identify the total number of samples, as well as the number of each sample's features.

The total number of samples is 350085, with each sample corresponding to 67 features.

Most of the features are float types, i.e. numerical data. However, nominal features are also present:

- iso_code: a string corresponding to each country's code.
- location: a string corresponding to each location's name.
- continent: a string corresponding to the continent where the location belongs.
- tests_units: a string corresponding to the units used in each location in order to count the number of tests (more details below).

There's also the date feature, the type of which is string, however it will be properly transformed into a datetime object in what follows.

	iso_code	location	continent	date	tests_units
0	AFG	Afghanistan	Asia	2020-01-03	NaN
1	AFG	Afghanistan	Asia	2020-01-04	NaN
2	AFG	Afghanistan	Asia	2020-01-05	NaN
3	AFG	Afghanistan	Asia	2020-01-06	NaN
4	AFG	Afghanistan	Asia	2020-01-07	NaN

It was observed that the tests_units column contains a notable number of null values, suggesting their presence in the dataset. To delve deeper into this observation, we can utilize the Pandas library to analyze the extent of null values in our dataset. By employing Pandas functions, we can efficiently count the exact number of null values for each feature. The outcome will be organized and presented in a systematic manner, offering insights into the completeness of our data across different attributes.

	0
weekly_icu_admissions	339880
weekly_icu_admissions_per_million	339880
excess_mortality_cumulative_absolute	337901
excess_mortality_cumulative	337901
excess_mortality	337901
...	...
new_deaths_per_million	9574
iso_code	0
location	0
date	0
population	0

3. Data Preprocessing

Given the evolving nature of the Our World in Data (OWiD) dataset on Covid-19, it is prudent to narrow our analysis to a specific time interval to ensure the consistency and relevance of our findings. For this purpose, we will create a filtered version of the full DataFrame, focusing on data recorded between January 1, 2021, and February 28, 2021.

To facilitate this temporal filtering, we will first ensure that the date feature is appropriately transformed into a date type object. This step is essential for accurate temporal selection and sets the stage for subsequent analyses within the specified time frame. By restricting our examination to this particular period, we aim to provide a snapshot of the data that minimizes the impact of potential future updates, contributing to the stability and interpretability of our analysis.

The total number of samples is 14935, with each sample corresponding to 67 features.

3.1. Handling Missing Values

Even in this filtered version, there's a sizeable number of null values present. Before investigating how to deal with them, it's important that we understand the reason why they're missing. As far as the continent feature is concerned, the following command sheds light into the reason why it contains null values.

	iso_code	continent	location
344958	OWID_WRL	NaN	World
344959	OWID_WRL	NaN	World
344960	OWID_WRL	NaN	World
344961	OWID_WRL	NaN	World
344962	OWID_WRL	NaN	World

Clearly, OWiD have performed a series of aggregations based on criteria such as income, or general aggregations (for example on the continent level). Since they may prove to be useful later on, there is no reason to discard them. The null values can simply be set equal to the 'OWID' value, in order to be able to invoke them later on if we need to.

Another column which corresponds to a nominal feature with missing values is tests_units. The distinct values that this feature assumes are:

nan, 'tests performed', 'units unclear', 'samples tested', 'people tested'

In other words, tests_units is simply a variable that indicates how each country/location reports on the performed tests. For example, in the case of people tested, the reported number of total tests is expected to be lower compared to the same report in the case of tests performed, since one person

can be tested more than once during the same day. This implies that the missing values are due to some countries/locations not providing the relevant information on how they count the total number of daily tests. Of course, this is not a reason to discard the relevant data, therefore the missing values will be replaced by the string 'no info'.

Moving on to the quantitative features, most missing values are due to the fact that the relevant data were either not available during the studied time period for some locations, or were simply equal to zero. For example, there are 11935 missing values in the new_vaccinations column, which are either due to the fact that vaccines were not available in some locations, or due to the fact that these locations reported no vaccinations for specific dates. The best approach in this case is replacing all these values with 0. In the few cases where the missing values are not due to any of these two reasons, but due to wrong reports, bugs, or other reasons, we expect to find it out during their analysis and especially their visualization. In this case, we will be able to re-handle them or discard them completely.

3.2. Outlier Detection

Having discussed the case of missing values, perhaps it's a good idea to also discuss the case of outliers. Typically, the identification of outliers requires further analysis, such as visualizations, since it is not a trivial matter (in fact, more often than not it's a case of a supervised learning problem on its own). Furthermore, there are several types of outliers, such as global outliers or context-based outliers (i.e. points that are outliers only given a specific condition or context), which means that dealing with outliers in a universal manner is ill-advised. Nonetheless, if one chooses to do so, a systematic way to deal with outliers is based on interquartile range methods. The interquartile range, R, is defined as

$$R = Q_3 - Q_1$$

where Q_i is the i-th quartile. Every point for which the studied feature has a value higher than $Q_3 + \alpha R$ or lower than $Q_1 - \alpha R$ is classified as an outlier for this specific feature, where α is a scalar that defines a "decision boundary" in units of R. This is essentially how Box plots are constructed, where R corresponds to the Box's height and αR is equal to the whiskers' length. One very common choice for α is $\alpha = 1.5$.

Based on these, one can define a function that identifies all outliers with respect to specific features.

For example, we can check if any of 5 random DataFrame rows correspond to outliers with respect to the new_cases feature:

```
     new_cases_outlier
364              False
365              False
366              False
367              False
368              False
```

3.3. Duplicate Entries

Before proceeding to the exploratory data analysis, the final step of the preprocessing phase is to locate possible duplicate entries and discard the duplicates. When speaking of duplicates we do not actually refer to a whole row, but rather the combined entries of the date **and** location columns. A duplicate entry on both of these features would imply that the location has provided more than one daily report on a given date.

4. Exploratory Data Analysis

Before diving into the EDA, we import some libraries and also present some helper functions and commands that will be utilized further down the road for visualizations.

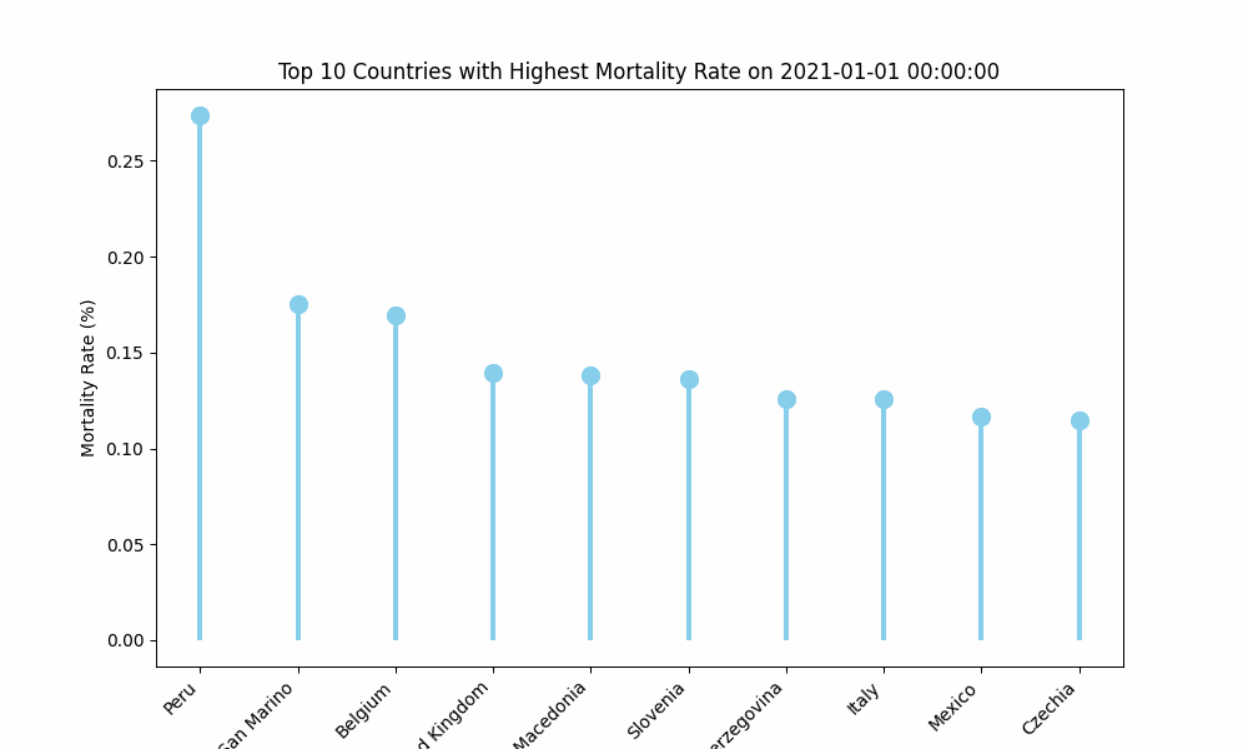
- **Evolution of top countries with respect to mortality**

Herein, the mortality rate is calculated as the total number of deaths divided by each location's population (another common definition is the total number of deaths by Covid divided by the total number of Covid cases). For this purpose, a column named mortality is constructed. Using this column, we identify the top 10 countries in terms of mortality rates, for every day of the studied time interval.

During 2021-01-01 00:00:00, the top 10 countries with the highest mortality rate were:
▶ Peru, with mortality rate 0.27%.
▶ San Marino, with mortality rate 0.18%.
▶ Belgium, with mortality rate 0.17%.
▶ United Kingdom, with mortality rate 0.14%.
▶ North Macedonia, with mortality rate 0.14%.
▶ Slovenia, with mortality rate 0.14%.
▶ Bosnia and Herzegovina, with mortality rate 0.13%.
▶ Italy, with mortality rate 0.13%.
▶ Mexico, with mortality rate 0.12%.
▶ Czechia, with mortality rate 0.11%.
This was the top ten until 2021-01-02 00:00:00, when United Kingdom, North Macedonia, Bosnia and Herzeg ovina, Mexico, Belgium, Czechia, San Marino, Italy, Slovenia, Peru joined the list, replacing .
This was the top ten until 2021-01-05 00:00:00, when Liechtenstein joined the list, replacing Mexico.
This was the top ten until 2021-01-14 00:00:00, when Mexico joined the list, replacing Liechtenstein.
This was the top ten until 2021-01-20 00:00:00, when Gibraltar joined the list, replacing Mexico.
This was the top ten until 2021-01-30 00:00:00, when Mexico joined the list, replacing Bosnia and Herzegovin a.
During 2021-02-28 00:00:00, the top 10 countries with the highest mortality rate were:
▶ Peru, with mortality rate 0.36%.
▶ Gibraltar, with mortality rate 0.28%.
▶ San Marino, with mortality rate 0.22%.
▶ United Kingdom, with mortality rate 0.22%.
▶ Czechia, with mortality rate 0.20%.
▶ Slovenia, with mortality rate 0.20%.
▶ Belgium, with mortality rate 0.19%.
▶ North Macedonia, with mortality rate 0.17%.
▶ Mexico, with mortality rate 0.17%.
▶ Italy, with mortality rate 0.17%.

The code systematically analyzes the mortality rates of the top 10 countries on each recorded date, presenting the data in an animated GIF format. This sequence of lollipop charts offers a visual narrative showcasing the fluctuations in mortality rates over time.

The resulting GIF, seamlessly merging individual charts, is shown below:



This lollipop chart not only provides insights into the countries with the highest mortality rates at different time points but also highlights the precise shifts in rankings among these countries. The dynamic visualization enhances the interpretability of the data, allowing for a more intuitive understanding of the trends in mortality rates.

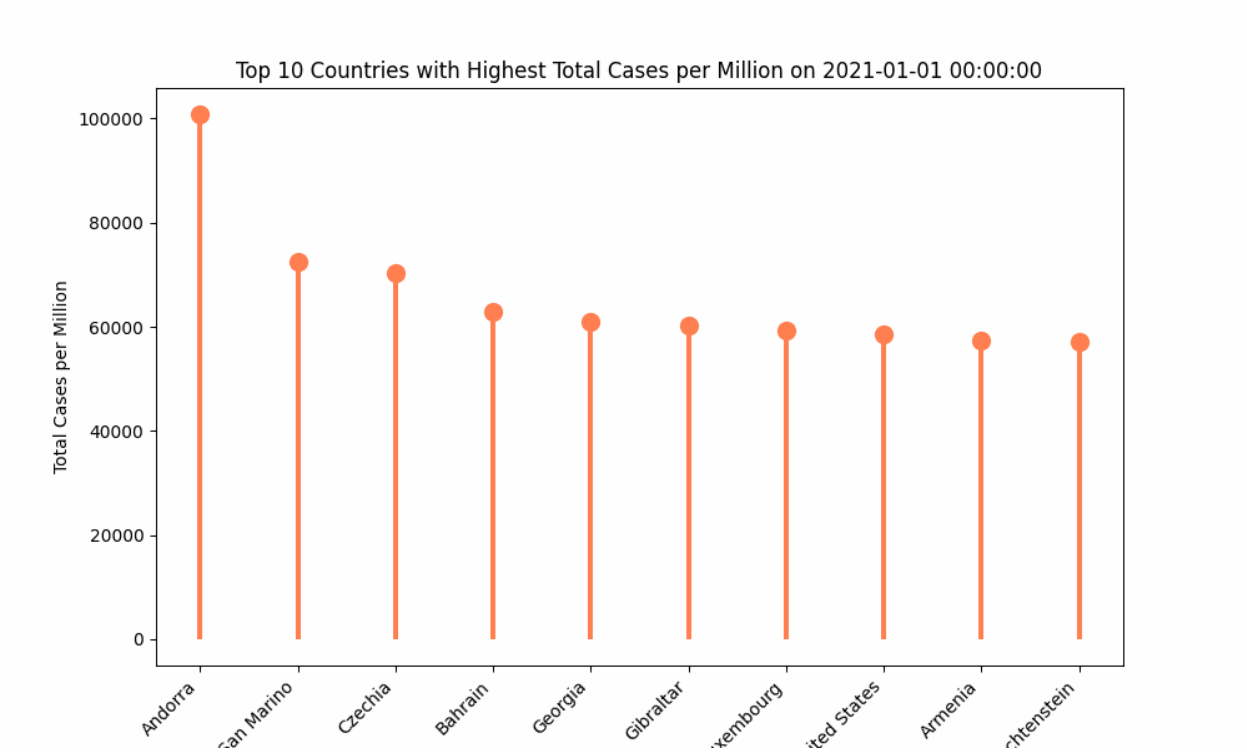
- **Evolution of top countries with respect to total cases per million**

The same procedure can be performed for the number of total cases per million. We choose to normalize the total number of cases in this way in order to be able to compare locations with different populations.

During 2021-01-01 00:00:00, the top 10 countries with the highest number of total cases per million were:
▶ Andorra, with 100810.34 total cases per million.
▶ San Marino, with 72573.464 total cases per million.
▶ Czechia, with 70230.675 total cases per million.
▶ Bahrain, with 62948.425 total cases per million.
▶ Georgia, with 61000.672 total cases per million.
▶ Gibraltar, with 60134.039 total cases per million.
▶ Luxembourg, with 59269.519 total cases per million.

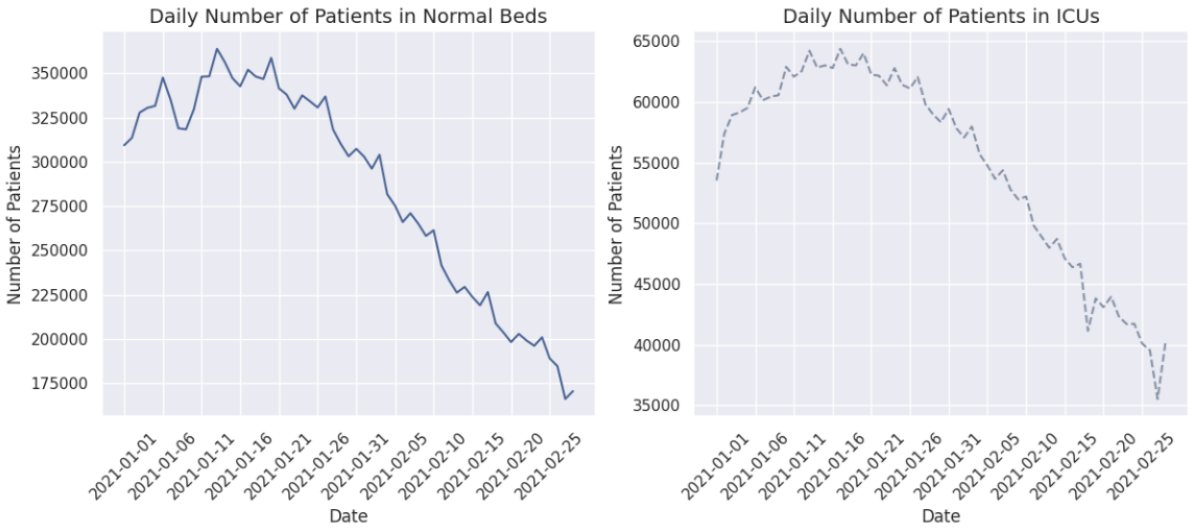
▶ United States, with 58569.309 total cases per million.
▶ Armenia, with 57449.958 total cases per million.
▶ Liechtenstein, with 57095.668 total cases per million.
This was the top ten until 2021-01-02 00:00:00, when Luxembourg, United States, Liechtenstein, Georgia, Armenia, Bahrain, Andorra joined the list, replacing United Kingdom, North Macedonia, Mexico, Belgium, Italy, Slovenia, Peru.
This was the top ten until 2021-01-03 00:00:00, when Slovenia, Montenegro joined the list, replacing Armenia, Liechtenstein.
This was the top ten until 2021-01-06 00:00:00, when Liechtenstein joined the list, replacing Slovenia.
This was the top ten until 2021-01-09 00:00:00, when Panama joined the list, replacing Montenegro.
This was the top ten until 2021-01-10 00:00:00, when Slovenia, Montenegro joined the list, replacing Liechtenstein, Panama.
This was the top ten until 2021-01-11 00:00:00, when Panama joined the list, replacing Luxembourg.
This was the top ten until 2021-02-02 00:00:00, when Israel joined the list, replacing Georgia.
During 2021-02-28 00:00:00, the top 10 countries with the highest number of total cases per million were:
▶ Andorra, with 135879.163 total cases per million.
▶ Gibraltar, with 129387.643 total cases per million.
▶ Montenegro, with 118958.924 total cases per million.
▶ Czechia, with 118903.963 total cases per million.
▶ San Marino, with 111694.865 total cases per million.
▶ Slovenia, with 89829.766 total cases per million.
▶ United States, with 83745.021 total cases per million.
▶ Bahrain, with 82716.302 total cases per million.
▶ Israel, with 81936.607 total cases per million.
▶ Panama, with 77072.628 total cases per million.

he corresponding .gif image can be seen below.



- **Hospitalized Patients and ICU Admissions**

Moving on, we study the hosp_patients and icu_patients features by visualizing the corresponding timeseries for the total number of hospitalized and ICU patients on a global scale.



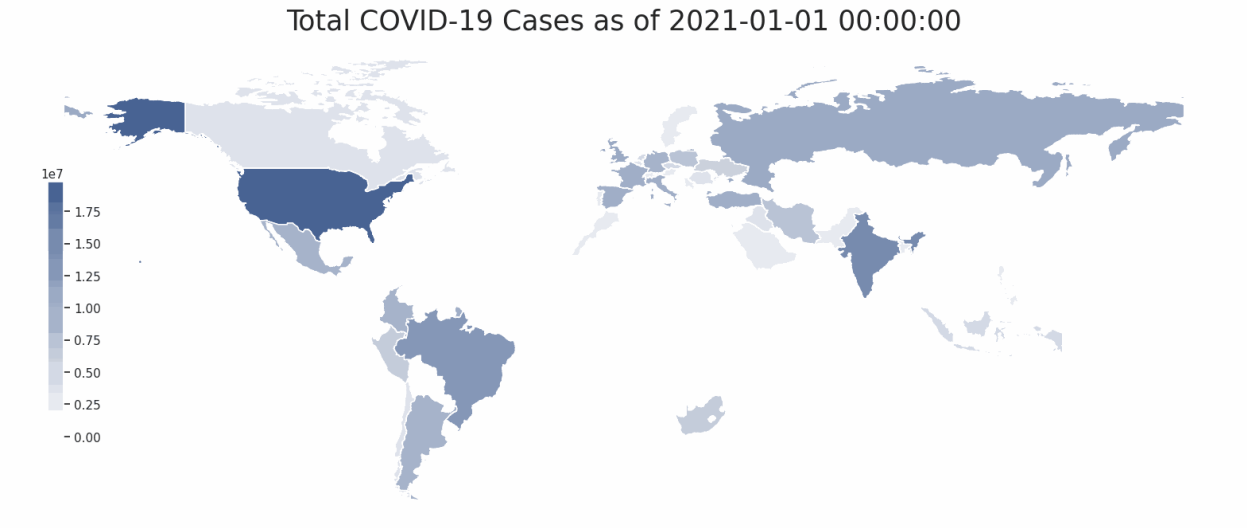
It becomes evident that the overall trend is downwards for both hospital and ICU admissions, since both numbers have declined to almost half their initial value by the end of the two-month period under study. It is worth noting that an upwards trend seems to appear near the end of February in the case of ICU patients. Of course, without further information we can't know if it is the beginning of a monotonically increasing trend, or simply a momentary increase, as the one identified between February 15-20. Finally, notice that both diagrams have a similar behavior, which hints at a correlation between the number of hospital patients and the number of ICU patients (which is probably expected). An important difference is that the absolute value of the number of hospital patients is considerably higher compared to the number of ICU admissions, which is reasonable, since the number of milder cases is higher compared to the number of more severe ones.

- **Geographic Heatmap of Total Cases**

An interesting visualization is the geographic heatmap, which is a 2D representation of countries world-wide which are colored depending on their intensity as far as a specific feature is concerned. Below, we construct the geographic heatmap for the number of total cases on a global scale. A heatmap image is extracted for each day and afterwards all images are merged into a .gif file. The heatmap is constructed using the geopandas library, as seen below. Note that to do this, we must first download a shapefile (.shp) which is the foundation for the construction of the heatmap and can be found here.

Construct heatmaps for every day

The final .gif can be seen below.



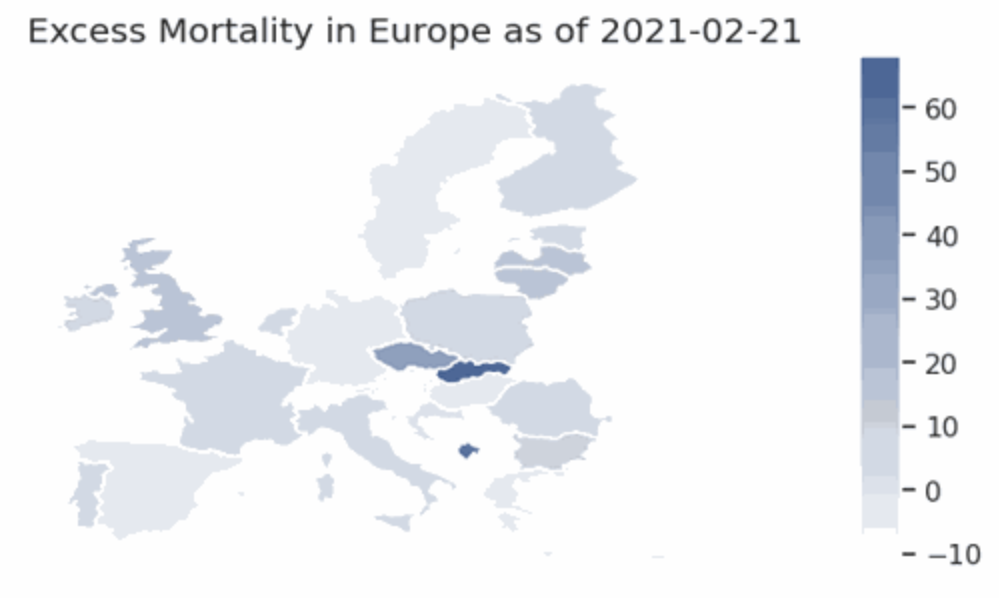
- **Geographic Correlation of Excess Mortality**

Based on the previous visualization it appears that some neighbouring countries are correlated with respect to the total number of cases (for example France and Germany). A reasonable hypothesis is that the same may be true for other features as well, such as the excess mortality.

The excess mortality is a feature for which the reports are weekly and not daily. It is equal to the total number of deaths for a specific week minus the mean number of deaths, based on reports from previous years. While it is not a feature directly connected with Covid, it's expected that during a global pandemic the excess mortality can be mainly attributed to this pandemic.

In order to investigate the correlation between neighbouring countries, we must first develop a list of dates for which reports on excess mortality are available (for all other dates, the entries are equal to zero due to our preprocessing).

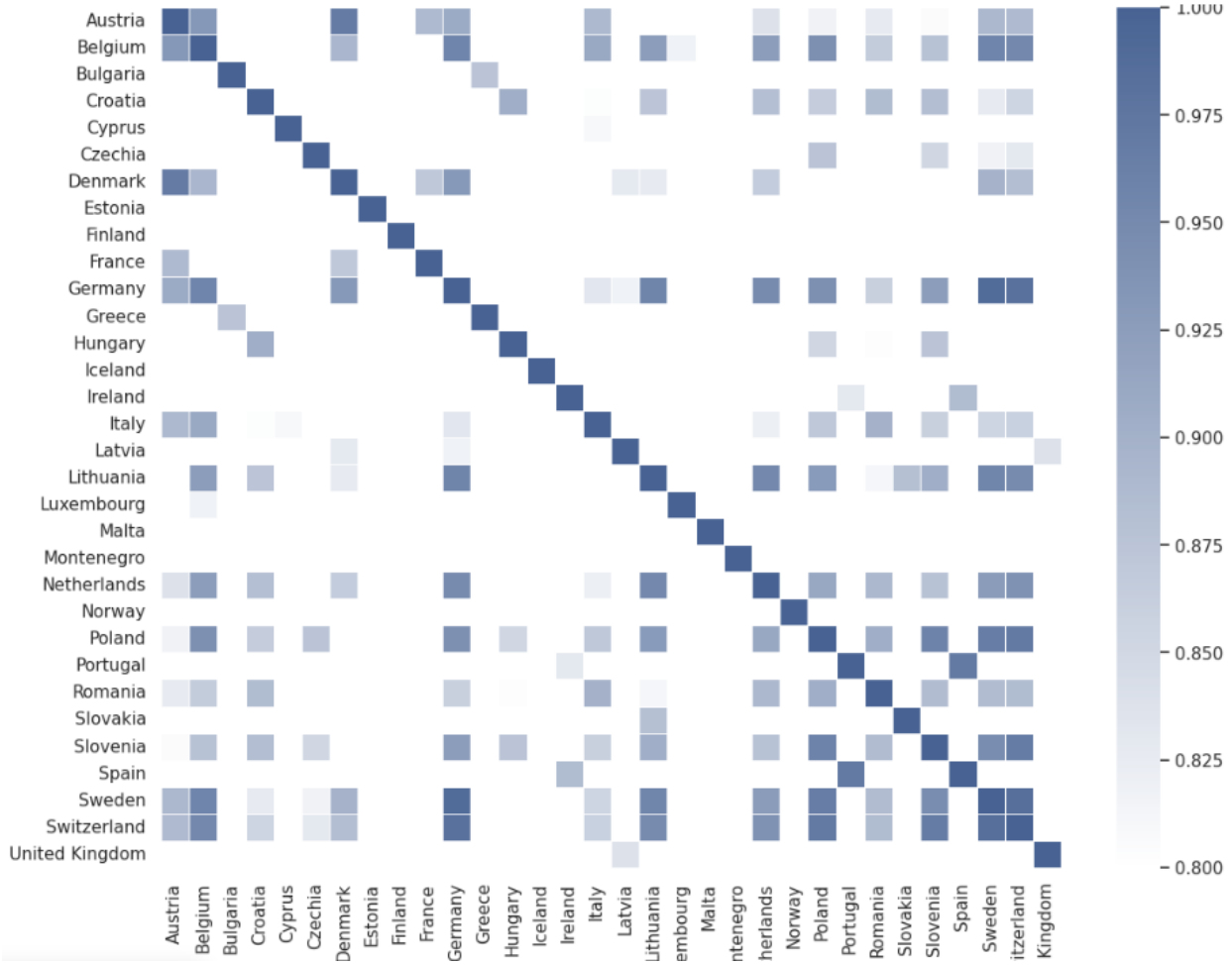
For brevity, we shall focus our study only on European countries. First, we construct a geographic heatmap of Europe with respect to excess mortality for each date calculated in the previous cell and merge the results into a .gif file, as done previously.



Based on this visualization, it's safe to assume that there indeed are neighbouring countries for which the excess mortality values appear to be significantly correlated. Germany and Switzerland are an example of one such pair of countries, as they appear to have highs and lows with respect to excess mortality at the same time.

In order to produce these results with more mathematical rigor, we need to construct a new PySpark DataFrame including all the reports on excess mortality for each European country that has provided reports on **all** of the previously calculated dates. Countries with even 1 missing value will not be taken into consideration, in order to be able to draw conclusions that are as safe as possible, since the volume of the available data is very small with regards to this feature. Then, using this newly created DataFrame, a Pearson correlation matrix can be constructed, thus revealing not only pairs of correlated countries that share the same geographical borders, but also the exact value of this correlation.

The calculated correlation matrix can be seen in the following heatmap, where only values of Pearson correlation that are higher than 0.8 are depicted (since we are looking for neighbouring countries with high correlation). This is why the lower limit of the colorbar is set to 0.8.



As noted before, Switzerland and Germany indeed correspond to a pair of highly correlated neighbouring countries with respect to excess mortality. In fact, it appears that in most cases **only** neighbouring countries (for example Belgium and Germany, or Luxemburg and Netherlands) and second neighbours thereof show high values of correlation, with the value of correlation declining significantly as the neighbour index (i.e. how many countries apart two countries are) increases beyond 2. Some additional examples of correlated pairs can be seen in the following table, along with the corresponding Pearson Correlation.

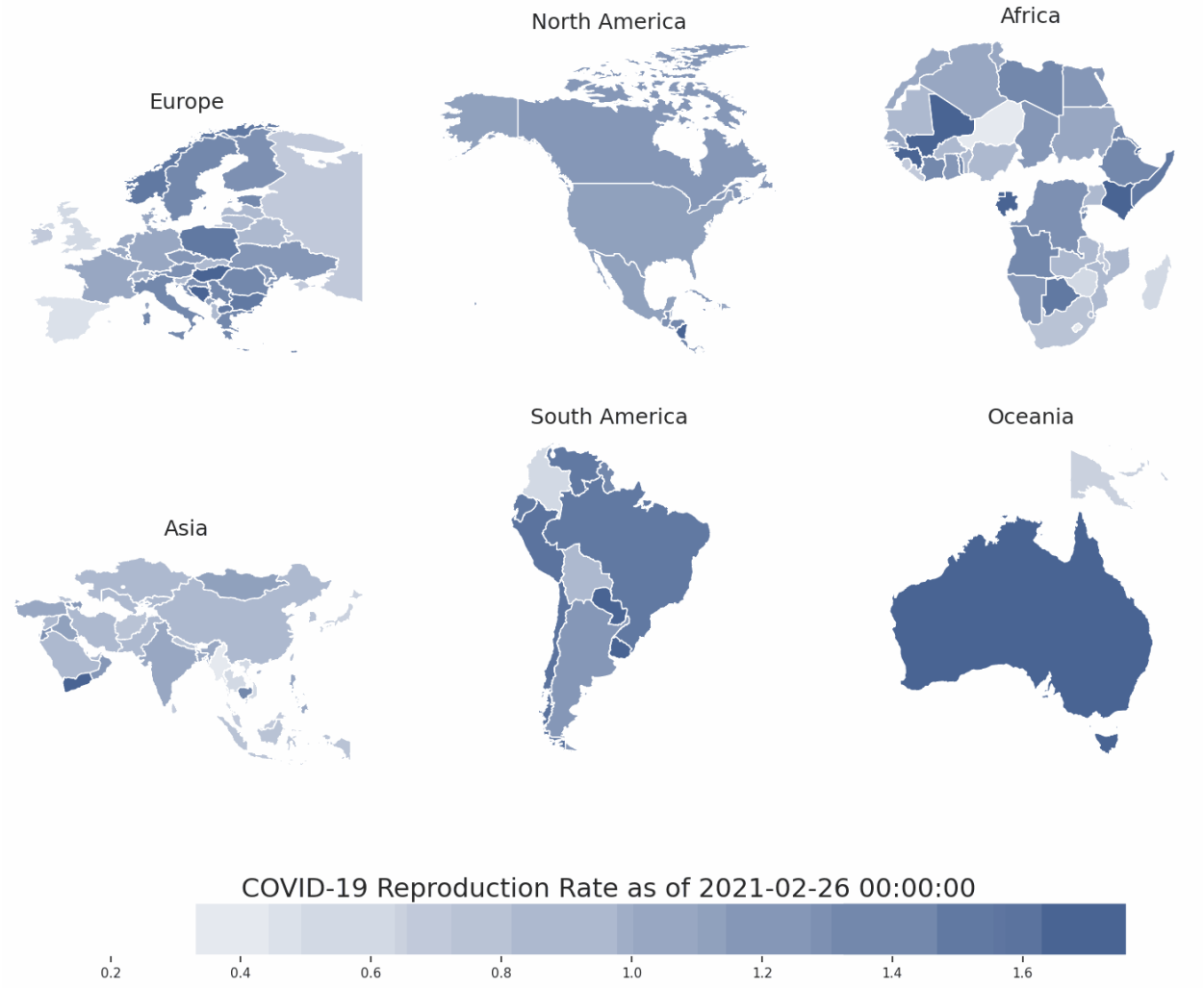
	Pearson Correlation for excess mortality (%)
Germany & Switzerland	98.19
Germany & Belgium	95.57
Lithuania & Poland	92.83
Netherlands & Belgium	92.60
Czechia & Poland	87.56
Italy & Slovenia	86.08

For a more extensive list of European countries with high correlation that is not limited only to neighbouring countries, one can run the following snippet of code:

```
for i in range(len(european_cts)):
    for j in range(i+1,len(european_cts)):
        corr_val = cor_np[i][j]
        if corr_val > 0.8:
            print(f'{european_cts[i]} and {european_cts[j]} show a correlation of {100*corr_val:.2f}.')
```

- **Reproduction Rate on the Continent Level**

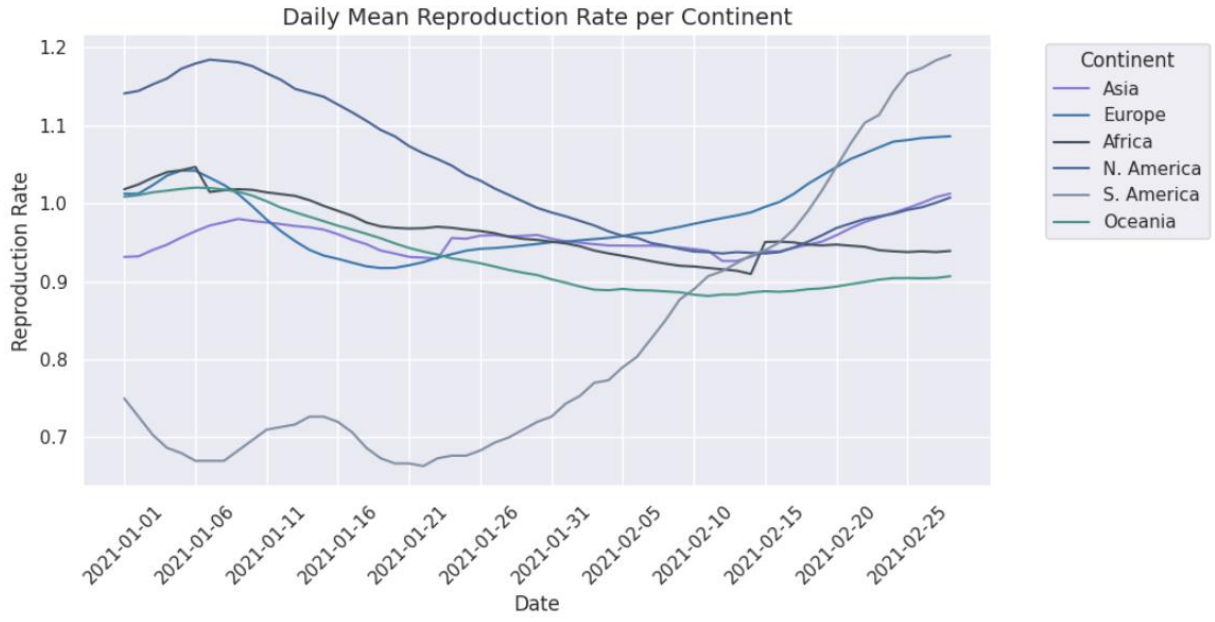
Moving on, we continue the study of the pandemic's features on a geographic viewpoint by grouping the countries together into continents. For this purpose, the DataFrame is split into continent-level DataFrames, in order to be able to draw the geographic heatmaps separately. The studied feature is now the daily reproduction rate, corresponding to the heatmap's intensity, with a common scale in order to be able to compare different continents. As usual, a .gif image is constructed using separate heatmap images for each day in the January-February interval.



On the continental level, it appears that Australia has both the highest and the lowest values for the reproduction rate. Such fluctuations are present in other continents as well, albeit less intense. In general, the reproduction rate appears higher during the last days of February compared to the first days of January, thus indicating an alarming trend of the pandemic. South America is a typical example of this trend, where the reproduction rate appears to stabilize on relatively high values as time passes. On the other hand, there's also the exception of North America (where the fluctuations are uniform for all its countries, thus indicating a high inter-country correlation), where the reproduction rate shows an overall decline during the two studied months.

On the country level, Asian countries tend to follow the pattern of neighbouring-countries correlation that was studied above for the case of excess mortality. In fact, in most cases this correlation can be seen between neighbours of order higher than 1, similar to what was observed in North American countries. Interestingly enough, in Europe, while Portugal and Spain show relatively high reproduction rates compared to other countries in the beginning of January, the exact opposite is true by the end of February. On the one hand, this observation can be attributed to the efficient crisis management by the two countries that gained experience after being struck hard by the pandemic during its early days.

The .gif image above depicts the reproduction rate on the country level as well, apart from the continent level. In order to focus solely on the continent level, we provide below the timeseries for the mean value of the virus' reproduction rate per continent.



The trends and fluctuations that were previously observed can be seen in this graph as well. For example, the overall trend towards higher values of reproduction rate is evident in all continents with the exception of North America. Nonetheless, an additional piece of information provided by this new graph is that during the last days of February the mean reproduction rate for the countries in North America has a tendency to increase. The same can be said for the countries of Asia and South America as well. Especially for the countries in South America, the overall increase of the

reproduction rate was approximately 60%, indicating that the overall increase may be even higher by the end of March, based on the aforementioned tendency. On the other hand, when it comes to Oceania, Africa and Europe, there is a tendency towards stabilization of the reproduction rate to a constant value.

Closing this part of the analysis, it's worth noting that the intense fluctuations that were observed for Oceania in the .gif above do not seem to appear in this graph, where the corresponding timeseries is close to being constant, if compared to the other ones. However, upon a closer inspection, it becomes clear that there is no inconsistency: the mean reproduction rate is calculated as an unweighted mean value by diving with the total number of countries, instead of a weighted mean with respect to each country's population. This means that the contribution of Australia is considered equal to that of Papua New Guinea for the calculation of the mean reproduction rate. As a result, the almost constant trend of the timeseries corresponding to Oceania can be attributed to the fact that the trends of Australia and Papua New Guinea as far as reproduction rate is concerned are inverse: whenever the rate is high for Australia, it is low for Papua New Guinea and vice versa.

- **Correlation Analysis between Excess Mortality and Daily Tests**

Shifting focus from geographic visualizations, our attention turns to exploring the correlation between different features at the country level. Specifically, we aim to understand the correlation between excess mortality and the number of daily tests performed. The key columns for this analysis are excess_mortality and new_tests_smoothed. The preference for new_tests_smoothed over new_tests stems from the former containing fewer missing values, enhancing the completeness of the analysis.

It's crucial to highlight that not all countries are considered in this analysis; rather, we focus on those with more than 5 **non-zero** entries for the excess_mortality feature. This filtering decision is rooted in the previously discussed challenges associated with this feature. Without this filter, numerous countries might exhibit extreme correlation values of +1 or -1 simply due to a sparse number of entries. For instance, Albania might show a correlation of +1 because it has only 2 entries for the excess_mortality feature during the months of January and February 2021.

As far as the correlation between new tests and excess mortality is concerned:

The ten countries with the highest correlation are:
South Africa, with correlation equal to 0.967.
Peru, with correlation equal to 0.962.
Portugal, with correlation equal to 0.959.
Spain, with correlation equal to 0.953.
Mexico, with correlation equal to 0.919.
Colombia, with correlation equal to 0.917.
United States, with correlation equal to 0.793.
Lithuania, with correlation equal to 0.777.
South Korea, with correlation equal to 0.718.
Ecuador, with correlation equal to 0.673.

The ten countries with the lowest correlation are:
Slovakia, with correlation equal to -0.883.
Denmark, with correlation equal to -0.820.
Romania, with correlation equal to -0.759.
New Zealand, with correlation equal to -0.722.
Switzerland, with correlation equal to -0.716.
Italy, with correlation equal to -0.698.
Sweden, with correlation equal to -0.643.
Luxembourg, with correlation equal to -0.536.
Guatemala, with correlation equal to -0.491.
Cyprus, with correlation equal to -0.413.

For the countries where the correlation is positive and close to one, elevated numbers of excess mortality seem to be related to an elevated number of daily tests and vice versa. This might concern countries which, by January or February 2021, had been severely affected by the pandemic and therefore performed a lot of daily tests, in order to be able to restrain the virus outbreak by isolating infected individuals and performing case tracking. [Spain and Portugal] are two examples of such countries.

On the other hand, in countries with the inverse correlation, either elevated numbers of excess mortality were not enough to pressure for more diagnostic tests (for example Romania), or despite low numbers of excess mortality, the daily tests performed were highly elevated for prevention (for example New Zealand or Luxemburg).

Before moving on, it is interesting to also study the correlation between the excess mortality and the course of the vaccination in each country, instead of the daily tests. Note that for this purpose we use the total_vaccinations feature instead of the new_vaccinations one, since vaccinations are a long-term measure. As a result, their efficiency cannot be imprinted on the number of daily vaccinations. For example, a country in which a high percentage of the population has been vaccinated (for example Israel) is expected to show small numbers of daily vaccinations, without this being an indication of an unvaccinated population.

As far as the correlation between excess mortality and the course of the vaccinations is concerned:

The ten countries with the highest correlation are:
Croatia, with correlation equal to 0.937.
Malta, with correlation equal to 0.713.
Peru, with correlation equal to 0.573.
Cyprus, with correlation equal to 0.444.
Chile, with correlation equal to 0.401.
Finland, with correlation equal to 0.379.
Ecuador, with correlation equal to 0.320.
Greece, with correlation equal to 0.252.
Bulgaria, with correlation equal to 0.204.
Australia, with correlation equal to 0.002.

The ten countries with the lowest correlation are:
Germany, with correlation equal to -0.986.
United States, with correlation equal to -0.985.
Switzerland, with correlation equal to -0.951.
Sweden, with correlation equal to -0.945.

Belgium, with correlation equal to -0.923.
Latvia, with correlation equal to -0.893.
Canada, with correlation equal to -0.879.
Poland, with correlation equal to -0.872.
Lithuania, with correlation equal to -0.852.
Italy, with correlation equal to -0.803.

In this case, most countries show a negative correlation (and more specifically close to -1), since the increase in total vaccinations is expected to lead to a reduction in excess mortality, as vaccinations have proven to prevent serious infections from Covid. However, there are still countries such as Cyprus or Croatia, where the correlation is positive. There, it's possible that vaccines became available for the general population during the studied time interval and as a result their efficiency on combating the pandemic has not yet been observed on large scales.

- **Covid and general health conditions on the country level**

Another interesting aspect of excess mortality is how it correlates with the general health conditions of a country's population. For this reason, we will first calculate the mean value of the 'female_smokers', 'male_smokers', 'diabetes_prevalence' and 'cardiovasc_death_rate' features, using the data on the last available date of our filtered DataFrame. Then, we will sort all countries with respect to their excess mortality per million, since a normalization is required when comparing different countries (and hence different populations). Finally, we will compare the values of the aforementioned features for the top 5 and the bottom 5 countries with their calculated mean values.

Based on data up to 2021-02-28 00:00:00, the mean percentage of female smokers is 10.79%, while the corresponding number for male smokers is 32.91%.
In addition, the mean percentage of people suffering from diabetes (aged 20-79) is 8.56%, while the mean number of deaths per 100.000 people due to cardiovascular conditions is 264.27.

The table below depicts the divergence of these characteristics from their mean values for each of the top 5 countries with respect to excess mortality (per million). The red color is used for the entries which are higher than their corresponding mean value.

	Female Smokers (%)	Male Smokers (%)	Diabetic Population (%)	Cardiovascular-related Deaths per 100.000 (%)
Armenia	-85.58	59.38	-12.87	29.17
Mexico	-33.65	-34.54	60.05	-42.14
Belarus	0.96	41.02	-36.52	67.82
Russia	125	78.34	-24.26	63.34
Albania	-31.73	56.62	23.53	15.20

First and foremost, it's evident that all of the above countries have at least one of the mentioned features assume a value higher than it's mean. With the exception of Mexico, the countries with the highest excess mortality are characterized by numbers of cardiovascular related deaths that are higher compared to their mean value. The same countries also show significantly increased percentages of male smokers, which is definitely correlated to the number of deaths due to cardiovascular causes. The percentages of female smokers do not show the same tendency, excluding Russia, where the percentage of the divergence from the mean is higher than 100%. Finally, as far as the diabetic population is concerned, Mexico (for which we observed lower numbers with respect to the other features, compared to their mean values) has a value higher than the mean value by 60%.

Moving on to the bottom 4 countries as far as excess mortality is concerned, it is worth noting that the negative values in these cases is not because of a bug or missing data. On the contrary, it is because the reported numbers of deaths from these countries during the studied period were lower compared to their expected number, based on previous years' reports. A table similar to the one shown for the top 4 can be seen below.

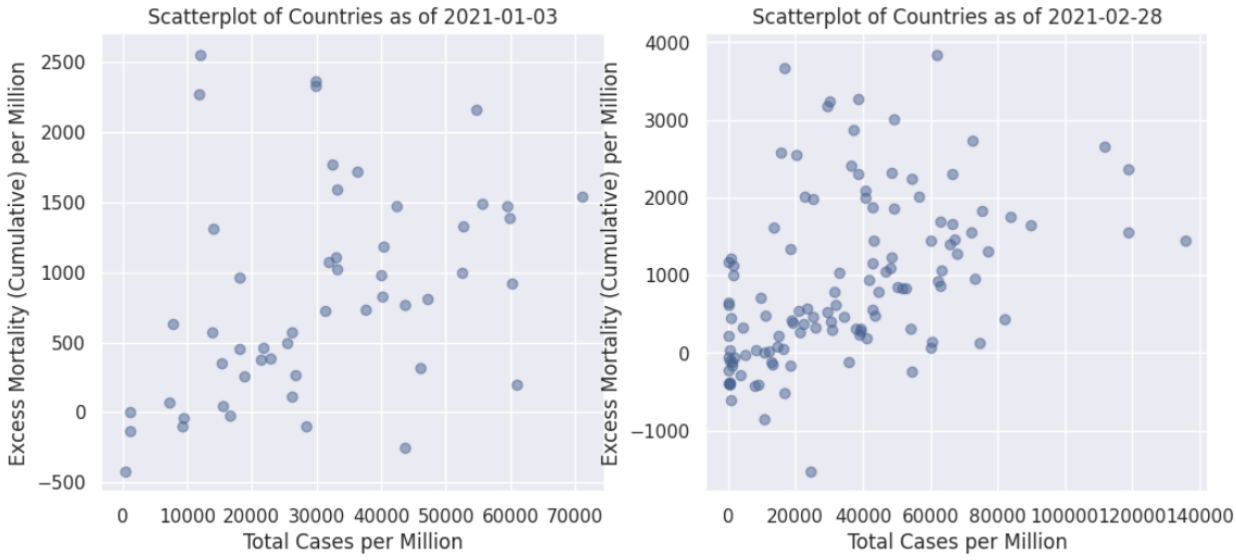
	Female Smokers (%)	Male Smokers (%)	Diabetic Population (%)	Cardiovascular-related Deaths per 100.000 (%)
Seychelles	-31.73	9.21	29.29	-8.11
Barbados	-81.73	-55.64	66.30	-35.60
Uruguay	34.62	-39.13	-15.07	-39.14
Mongolia	-47.12	42.25	-40.93	74.21

In this case as well, all countries have at least 1 studied feature with value higher than the corresponding mean. Nonetheless, the most important issue that was present for the top 4 countries, i.e. the elevated numbers of death by cardiovascular causes, does not seem to appear in this case as well, with the exception of Mongolia. When it comes to the percentage of smokers, it appears significantly reduced for both sexes. In contrast to what was observed for the top 4 countries, the countries with the lowest excess mortality tend to have increased diabetic populations.

Based on these, one could conclude that the excess mortality due to Covid can be connected to high percentages of smokers in the general population, as well as cardiovascular diseases. Similar conclusions cannot be drawn for the case of diabetes, which may be uncorrelated with deaths due to Covid. Of course, more extensive studies need to be performed in order to draw such conclusions safely, as well as more tests on target groups.

- **k-Means Clustering**

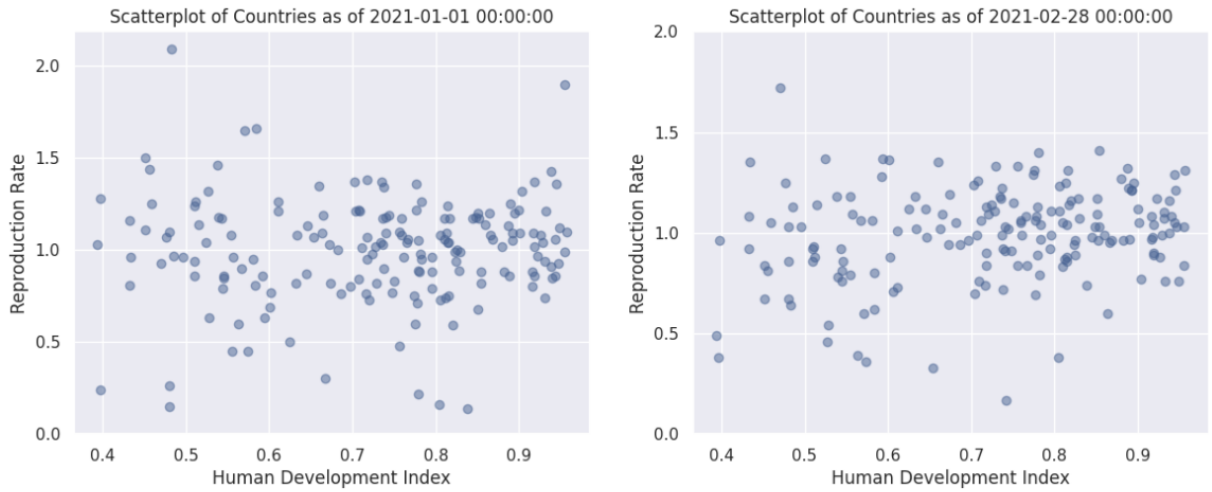
Moving on to the final part of this EDA, we incorporate unsupervised learning methods, and more specifically k-Means clustering, in order to draw some additional information from our data. This is our final study on excess mortality and we intend to cluster countries together with respect to it, as well as the total number of cases - both normalized. This clustering will be performed on two different dates: the first and the final date present in our filtered DataFrame, in order to be able to see the evolution of the initial state. As previously done, we will only take into account countries with no missing values (i.e. zeroes) on excess mortality.



Even through this preliminary visualization we can extract a very important conclusion as far as the data themselves are concerned: the number of countries which report on excess mortality has increased by the end of February, since the second scatterplot includes more points.

As far as the choice of k is concerned, i.e. the number of clusters to be taken into account, a reasonable hypothesis for the first date is k = 2: one cluster that includes the countries with fewer covid cases per million and one that includes the countries with more covid cases per million, since - with the exception of some outliers - it seems that the excess mortality is proportional to the number of total cases.

Closing our investigation on clustering and the project itself, we perform the same steps in order to cluster countries with respect to the virus' reproduction rate and the countries' human development index.



As illustrated in the animated GIF below, the scatterplots depict the time evolution of countries based on their human development index and reproduction rate. The motion is predominantly vertical, indicating minimal changes in the human development index over the observed period. Notably, during the initial days of January, some countries exhibit a reproduction rate exceeding 1.5. However, as time progresses, the oscillatory motion converges to the [0.5, 1.5] window for the reproduction rate.