Basics of Data Science Course - Assignment 2

Soroush heidary

96222031

Data Set : Covid-19

(For a more complete experience run the notebook as the visualizations are interactive)

Summary :

Most of the visualization we have used are animated over time so this report would just be a quick walkthrough over what has been done, and the main part would be the code for you guys to run and see for yourselves.

Also as the main reason of this exercise is to familiarize with different Visualizations and the respective applications we'll have some plots just for the sake of that purpose only. (e.g. sunburst plot) meaning a lot of the visualizations seen here are actually unnecessary for a concise and fruitful EDA

This is a list of what you will see in code, we'll have a simple output of each here just for introduction, again you need to run the code for more detail

- Spreading of the virus over the globe
- Continent wise comulative stacked plot over total_cases
- Continent wise percentage based plot over new_cases
- Chronological EDA on different time-stamps over new_cases/new_deaths
- Outlier detection and why we won't be doing that!
- Different choroplets over columns

From now on we'll use a collection of selected countries for more details

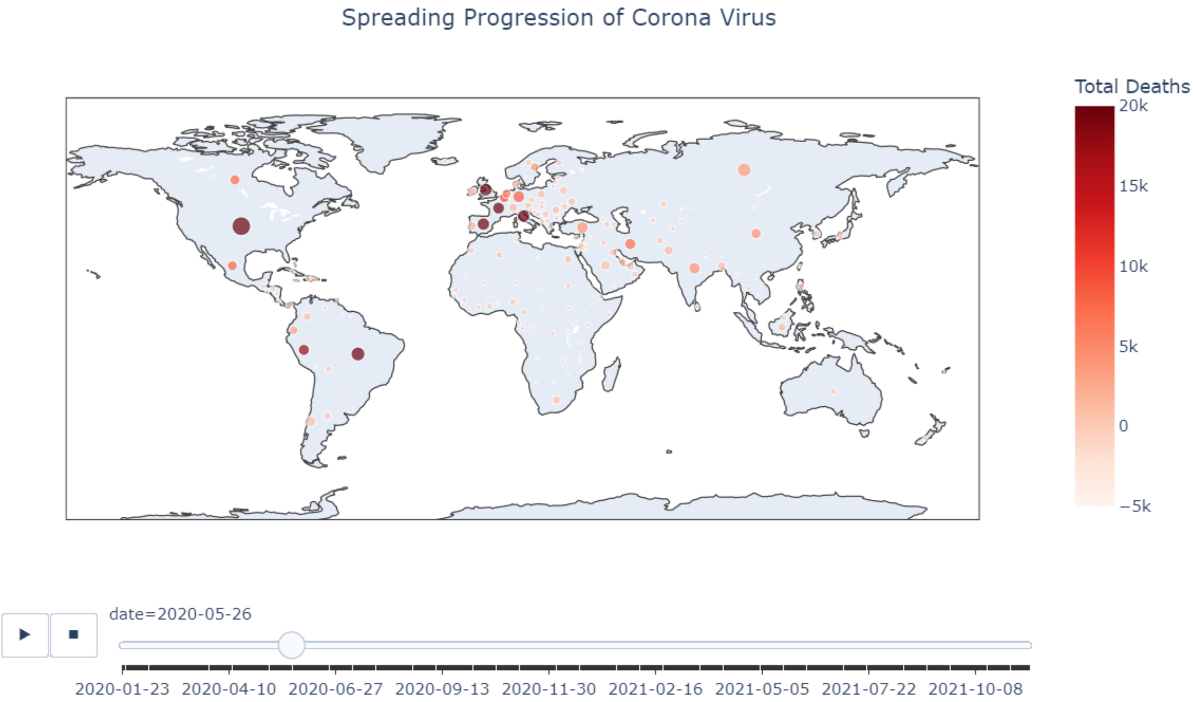
- Weekly differences of new cases
- Recovery rates of each country
- Mortality rate of each country
- Wave counting of countries + world wise

Country specefic features such as gpd_per_capita

- Heatmap of all correlations to eachother
- Parallel coords of a fraction of these features
- Sunburst and treemap plots

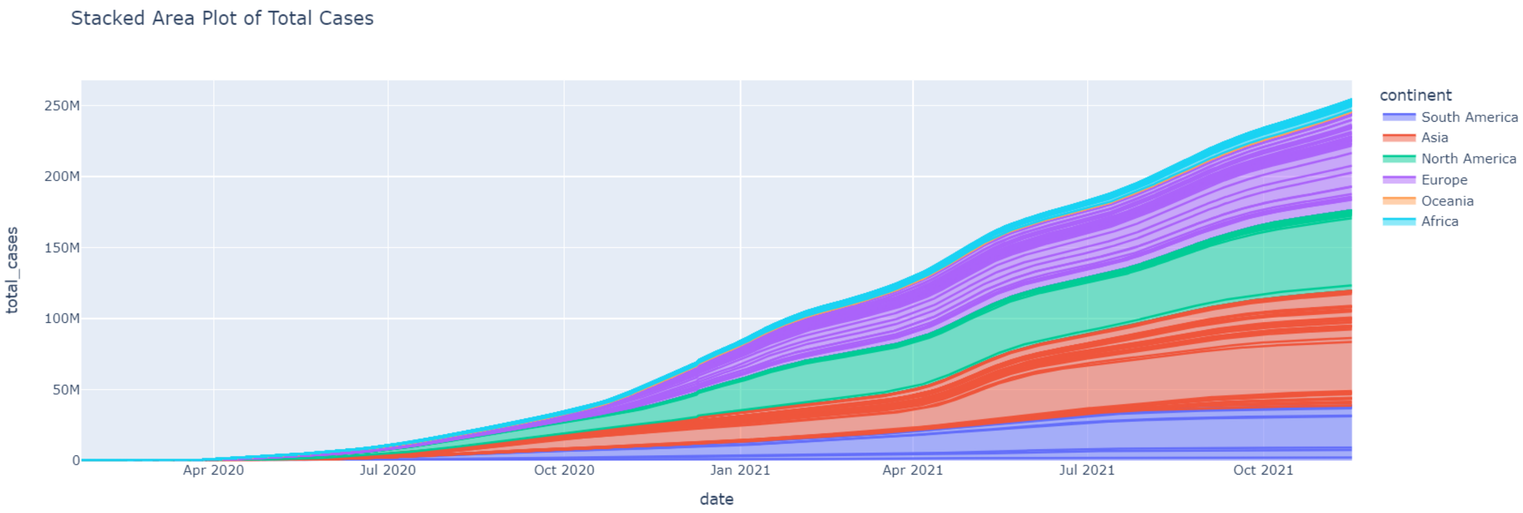
1- How the virus spreaded

Using a scatter plot on global map (with plotly.express.scater_geo) we can animate through time and watch how the infection started, it is colored over death counts and sized with confirmed cases count, we see it would start and china and just stops there, other countried would get infected with a high number of casualties but china remains the same after the first month of infection



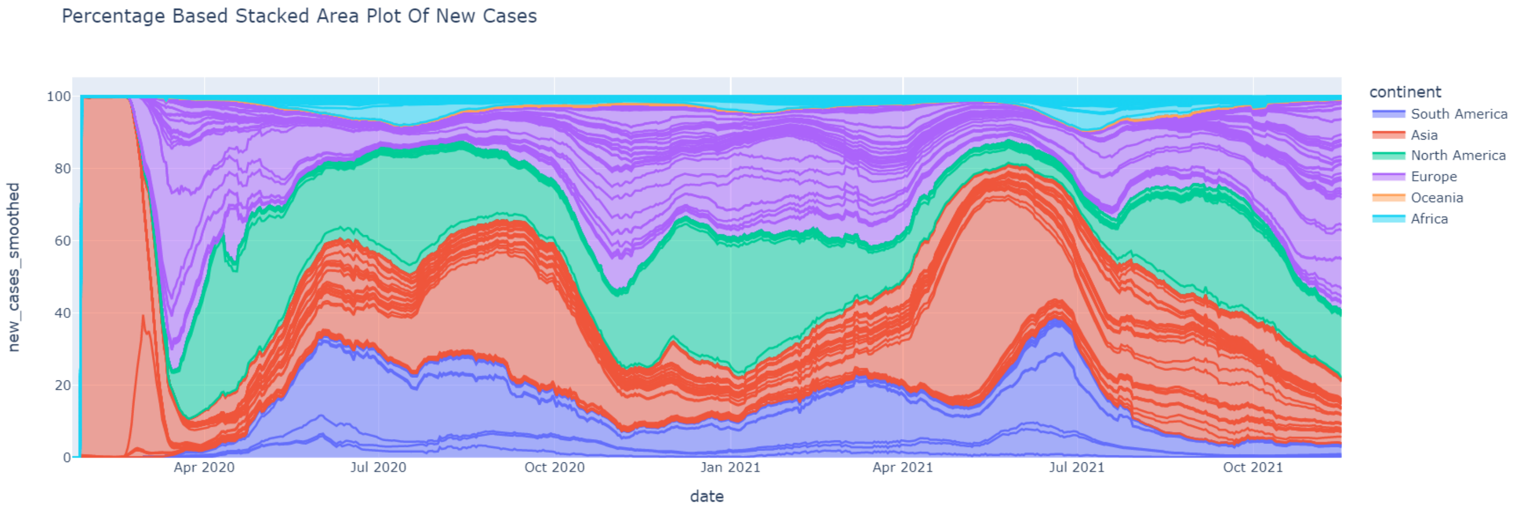
2- Continentwise comulative stacked area plot

Area plots are good for analyzing moments of progression let it be increasing or decreasing, using a stacked area plot let's us to compare each of the continent's total cases over time, the worldwide infecion slope would be the last stacked area's slope(here it would be Africa)



3- Percentage based stack area plot

In the above plot there's one thing we can't clearly see at times we have very low number of cases, using a percentage based plot solves that issue, now we can compare each contintent's new cases and clearly see in which time periods each of the continents were dominating ones in term of having more cases
the reasoning above is the same as when we use logarithmic scales when an outlier is overshadowing an slope



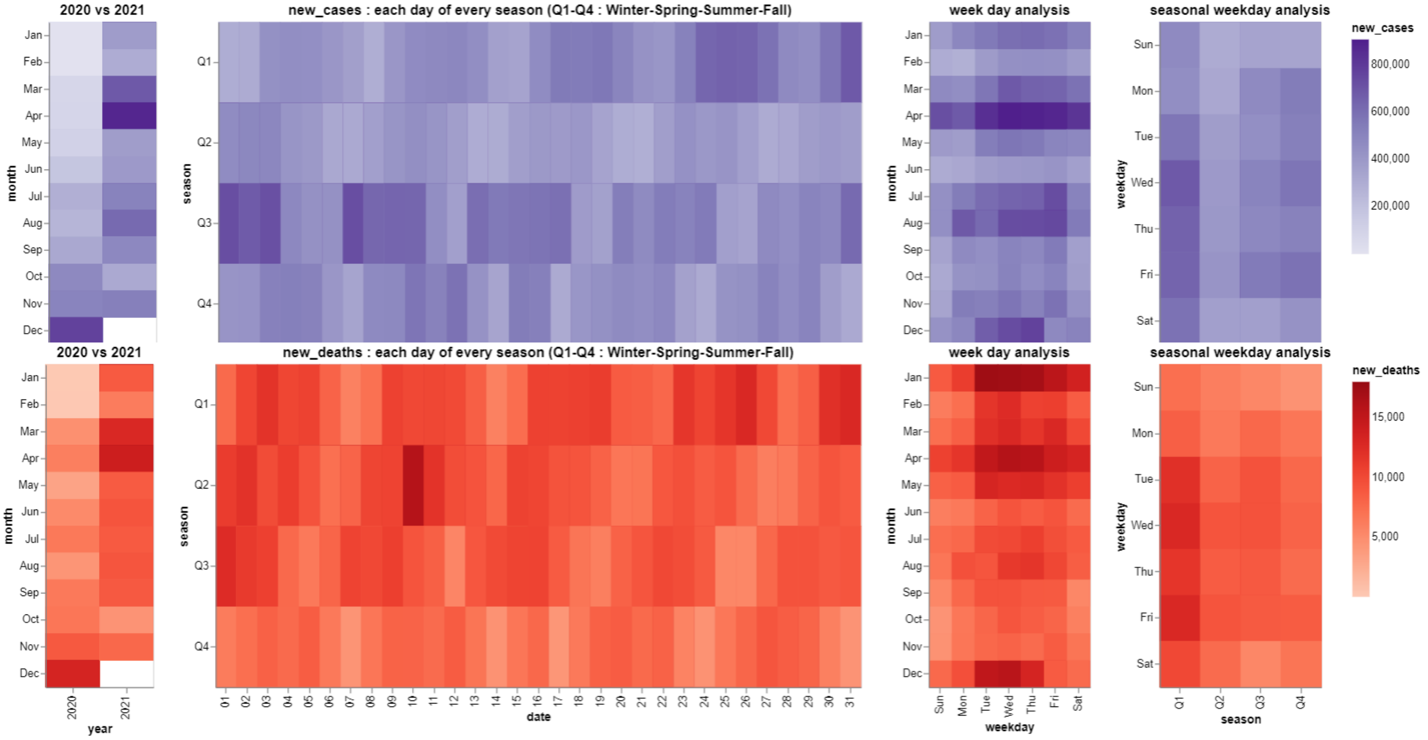
4- Chronological analysis on different time stamps

There are a few insights that could be derived from these heatmaps, each of them has a different time_stamp combination to help us derive different kinds of information

Some of these insights could be :

Week day distinguishment: from the two last chart we observe that as we reach the middle point of our week our cases and deaths increases in number an assumption for this observation could be that as we reach the weekend people tend to be out of their houses more than often thus causing new cases of infection to spread and show itself with a 2-3 day delay (the virus shows symptoms with a 2-3 day delay)

Season distinguishment: we have Q1 as Winter, and clearly has more death and cases submitted in this season, probably one of the main reasons is the tempreture, Spring is the lowest with cases/de



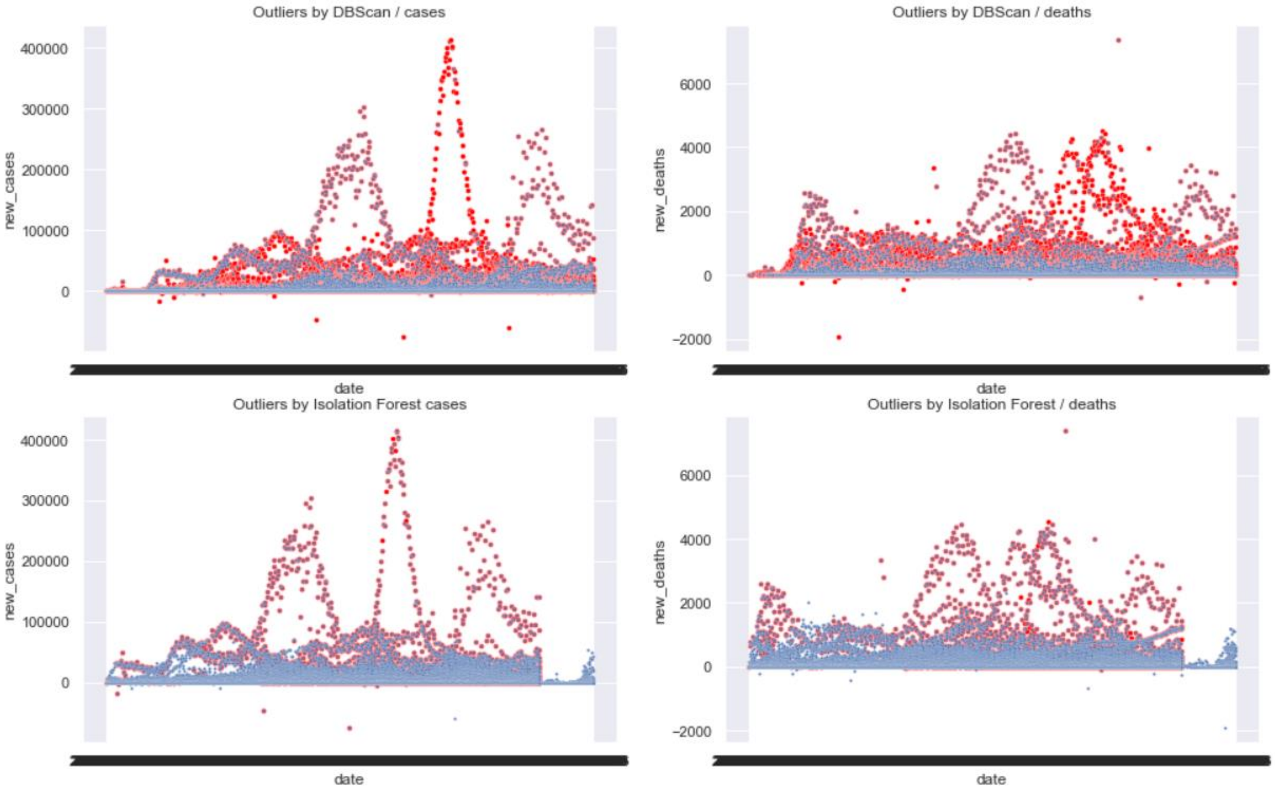
Before we delve into more detailed visualization we would want to know is there any outliers in our data (well ofc we wont be doing any modeling and we wouldnt need anomaly detection)

though as we see below the anomaly detection concept being a challenge in data science itself, there's not an easy path detecting outliers here, the main reason is that we have a lot of features to work with, but if we narrow them to only new/total cases/deaths which this happens :

5- Outlier detection

I have used two methods for detecting outliers including DBScan & Isolation Forest, both were trained over the whole dataframe as a n-dimentional input, but to visualize which of the rows were detected as outliers

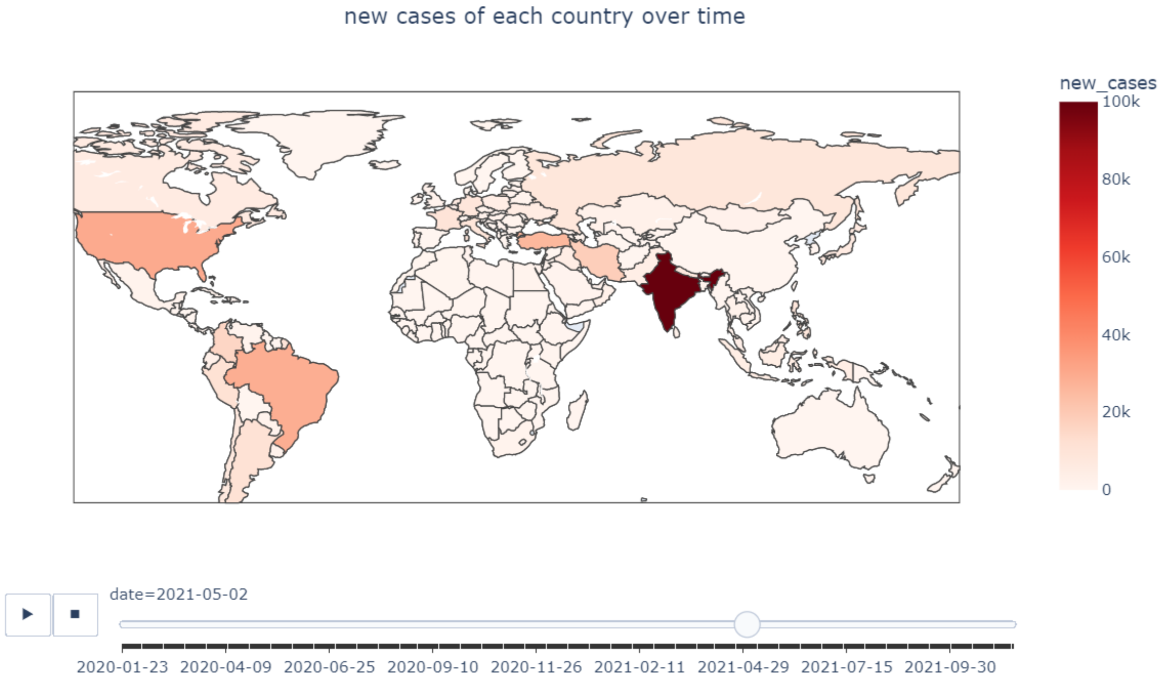
I have used new_cases(col1) and new_deaths(col2) and solely to show that removing outliers in this data set means removing all the infection waves which are an extremely important part of the data and we wouldn’t want to remove them, also we are not doing any modeling here so we don’t need to do anomaly detection either way!



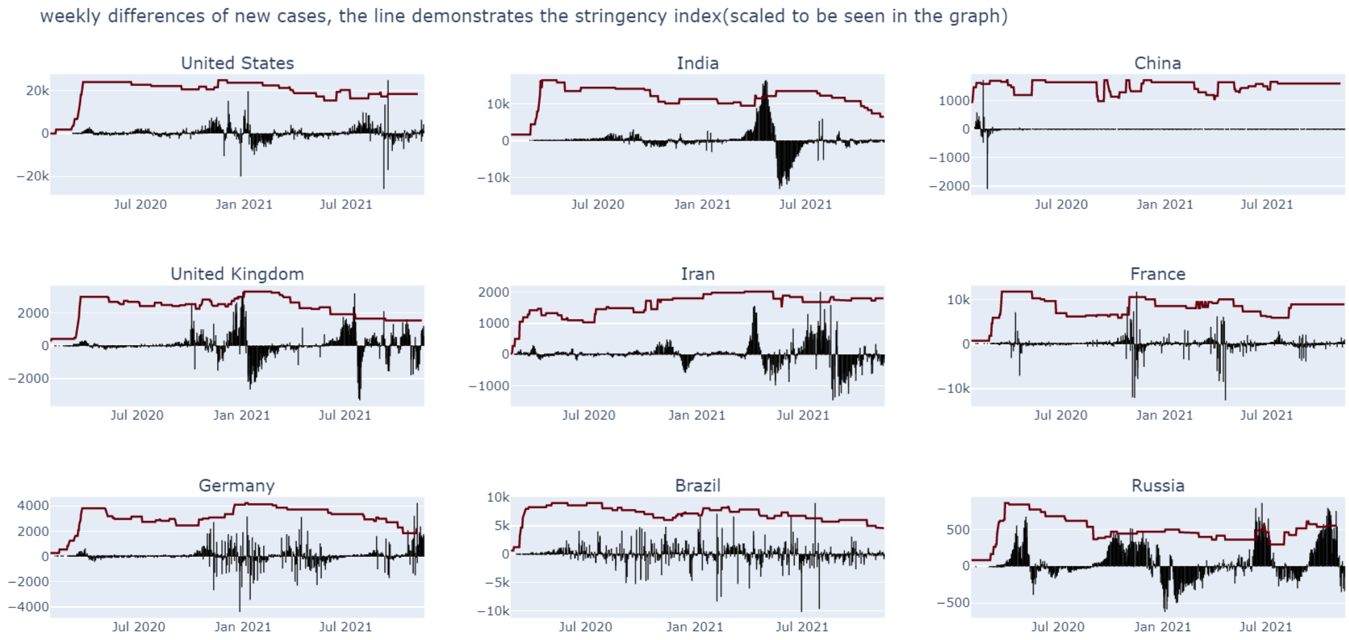
To have a more detailed visualization over our features for some cases we would want to pre select some countries to work with (except choroplots) these are: United States, India, China, United Kingdom, Iran, France, Germany, Brazil, Russia. How did I select these countries to be the our representative ones? Well just sheer geographical/political knowledge (I failed geography course in middle school (miserebaly))

Choropleths are good choices when we want to have an overall view of all the coutries but gives us only one dimention(the color) to work with, (although animating over a feature could be applied too, e.g. date) for this purpose we have a choropleth which shows the new_cases of each time_stamp by scrolling over it, I did try to make it more interactive with adding other features as a drop down button in the figure to select from, but that was a burden that made me wish for the apocolypse to happen right away, and failed to do so. Though viewing different features over the choropleth is just the matter of changing one name as the color method which requires you to do if you want to watch the other features as coloring factor.

6- Choropleths



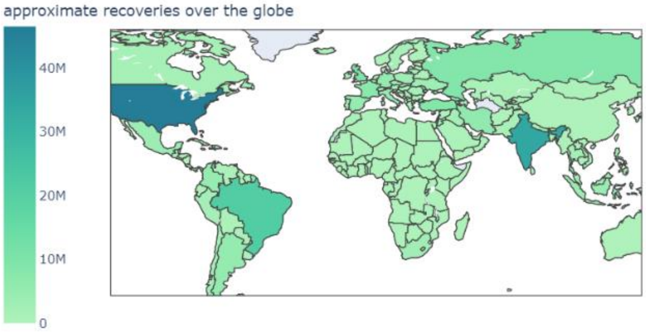
7-weekly diffs on new cases and str_index

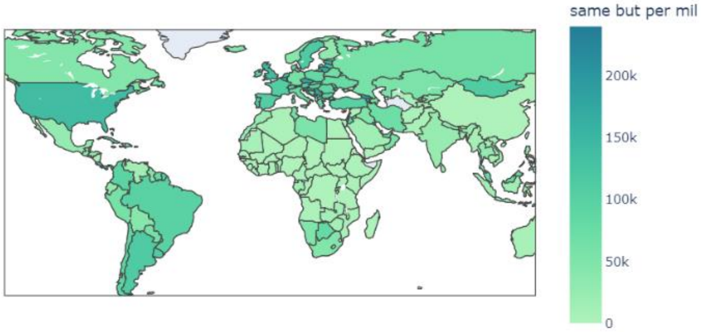


Something that would catch the eye is the stringency index, seemingly it shows how much a government got serious about the situation and took action to decrease the infection rate, actions like limiting the public gatherings or closing schools, or …
Here we see a chart of each of our selected countries which shows how each day's new cases were comparing to the last day (though smoothed over week to be more stable)
And on top of that we have our stringency index scaled to be seen in the graph

8 -recovs

Revocery number is not something we can confidently know, first we do not have the exact number of cases nor the exact number deaths, beside that we can't decide that how much time it took for each case to be cured so we can't animate this feature over time, though some tries took place to solve this issue which I explained in the code but at last I just used the total_cases – total_deaths (left fig) and total_cases_per_million – total_deaths_per_million (right fig)



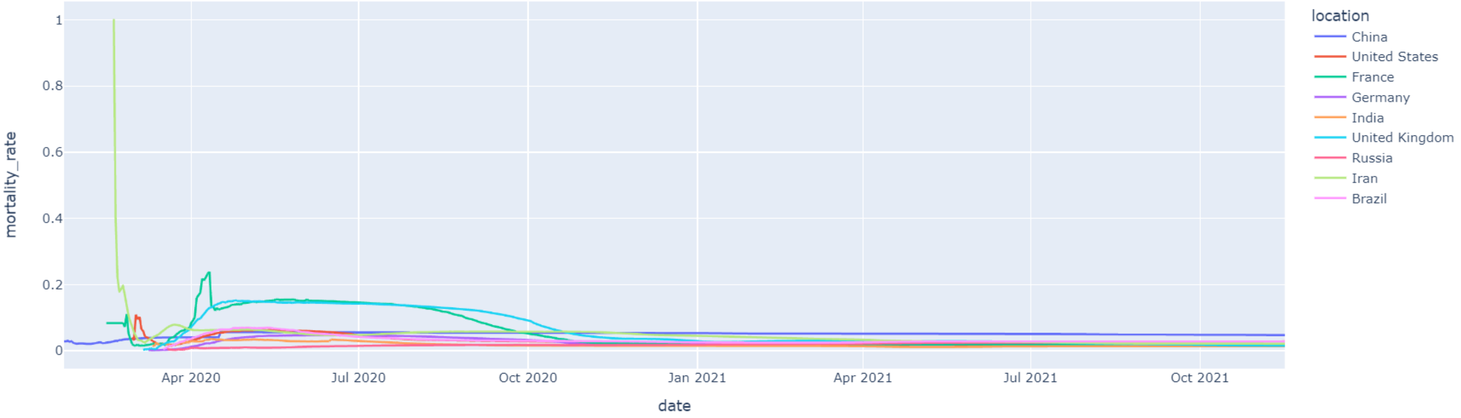


9- excess mortality and mortality rate

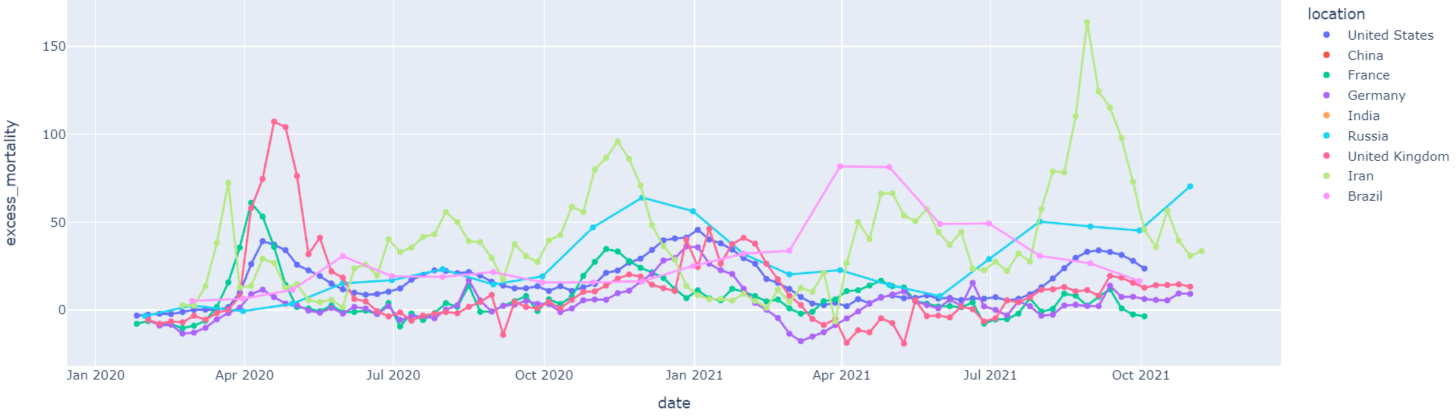
One more feature we could add is the death rate of new cases which is defined by the simple formula below :

df['mortality_rate'] = df['new_deaths'] / df['new_cases']

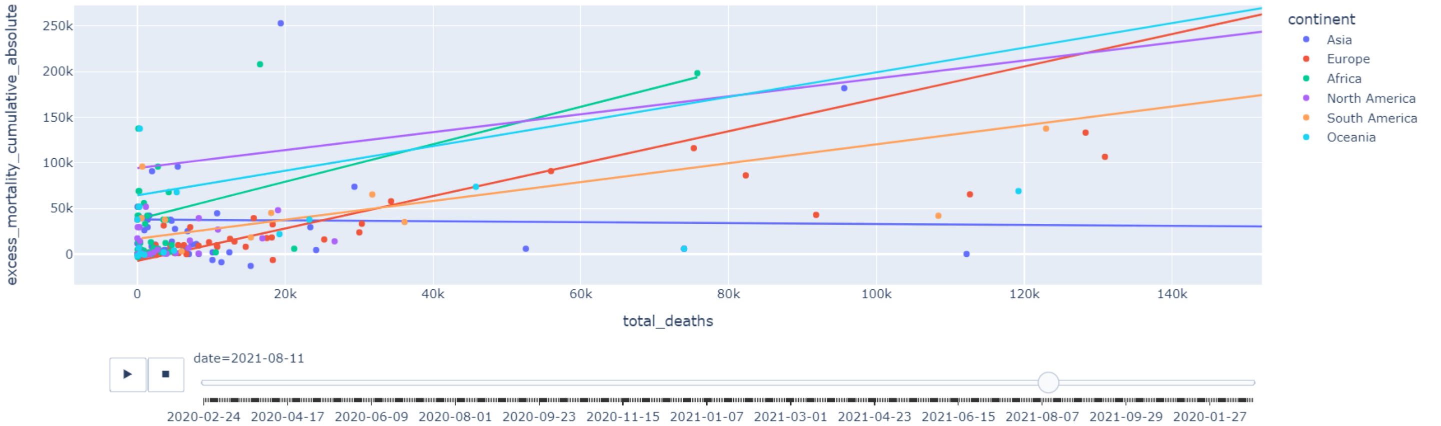
you'll see the above feature in the first figure, the following two are visualizations of Excess_mortality feature which is by efinition the number of deaths increased in a country comparing to the normal amount of that country, excess_mortality could be a good alternative to tota_deaths, as it is showing since the start of corona virus how was the death rate affected, meaning it is a summation of corona_deaths + other_corona_related_deaths mortality rate of each country over time



excess mortality of each country over time

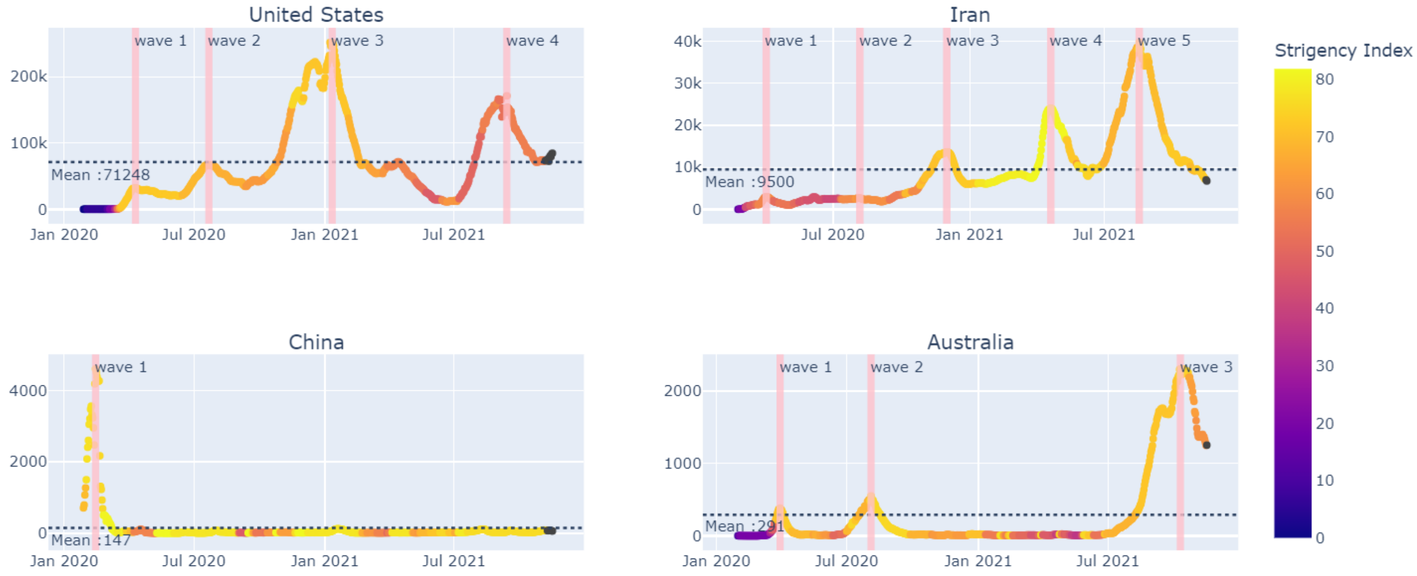


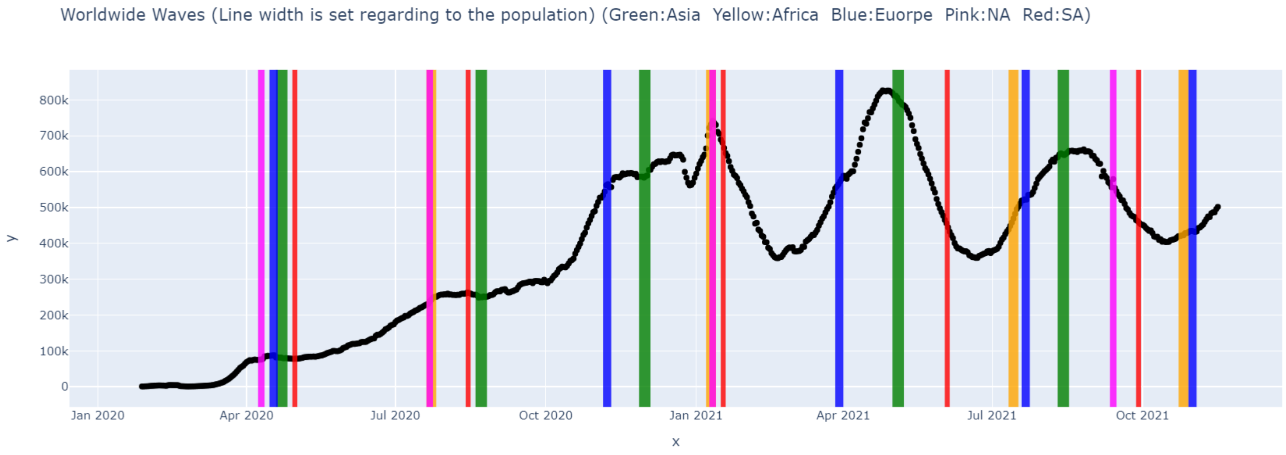
Excess Mortality vs Total Deaths



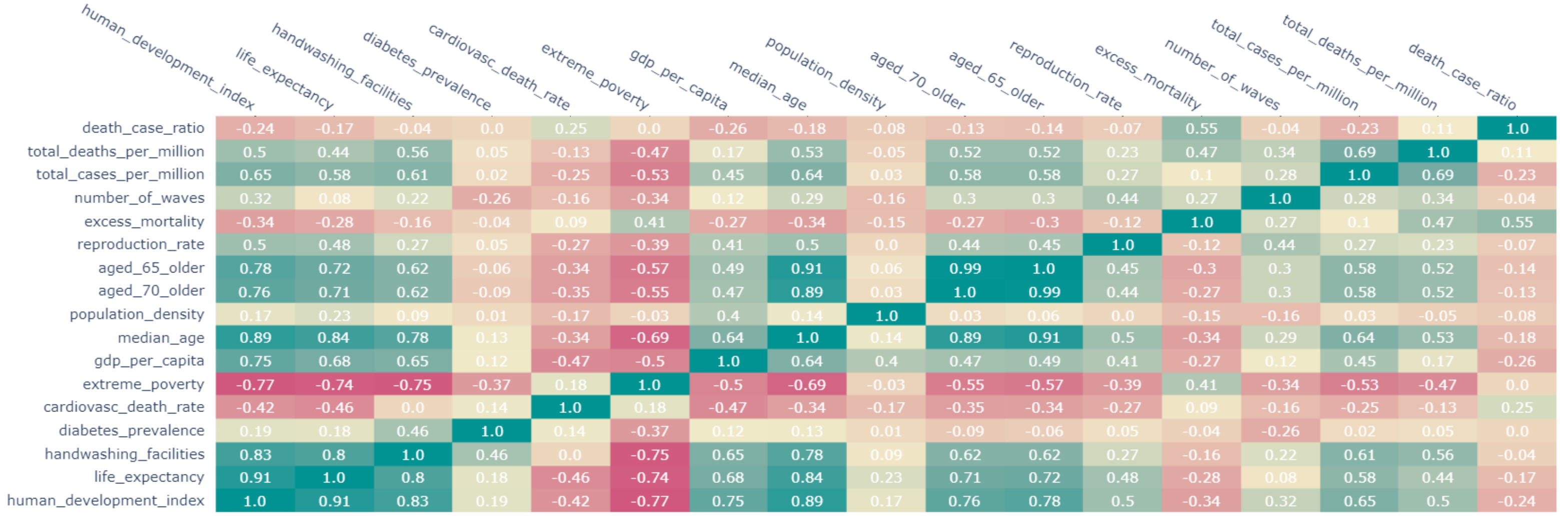
10 – wave counting

As I tried to add one more feature to our data set, which is the number of infection peaks over each country, for this purpose two algorithms where tested out (both written in sklearn) and after some tuning the model we the results for United States, Iran, China, Australia (what’s up with this china dude, srsly?) I couldn’t find a somewhat “usefull” applications for this feature though we see the continent wise waves affecting the world chart below (in the code I marked down some self-made analysis of this)





11- heatmap

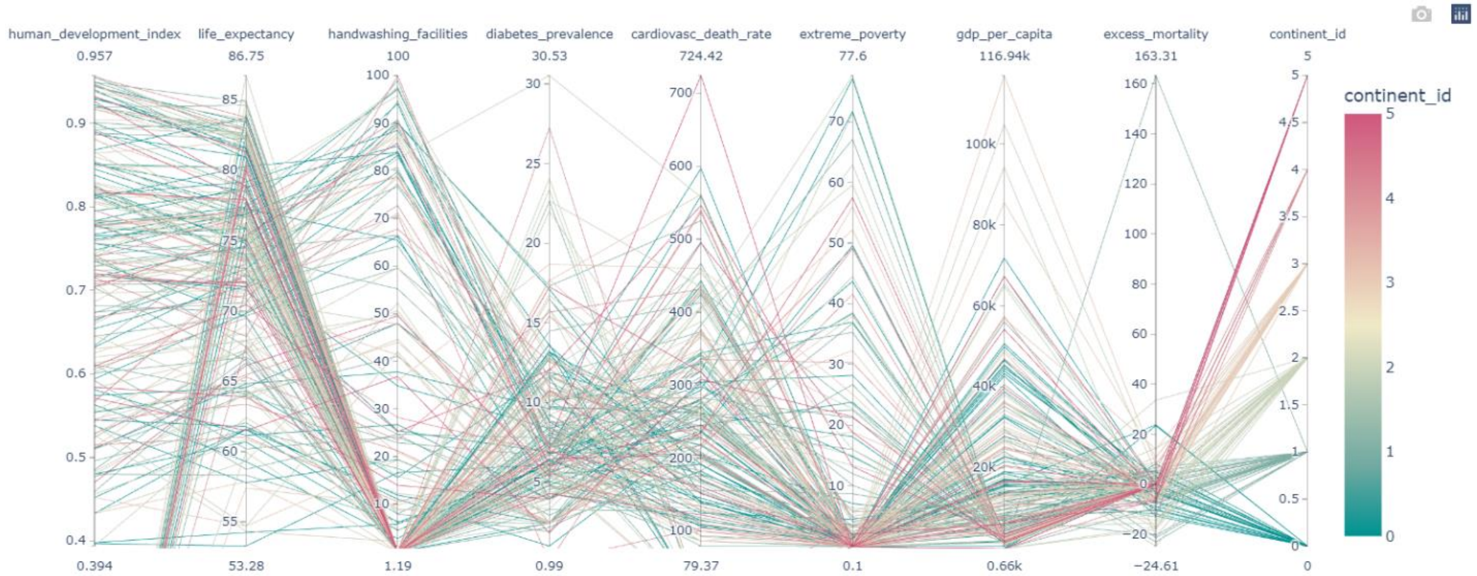


	human_development_index	life_expectancy	handwashing_facilities	diabetes_prevalence	cardiovasc_death_rate	extreme_poverty	gdp_per_capita	median_age	population_density	aged_70_older	aged_65_older	reproduction_rate	excess_mortality	number_of_waves	total_cases_per_million	total_deaths_per_million	death_case_ratio
death_case_ratio	-0.24	-0.17	-0.04	0.0	0.25	0.0	-0.26	-0.18	-0.08	-0.13	-0.14	-0.07	0.55	-0.04	-0.23	0.11	1.0
total_deaths_per_million	0.5	0.44	0.56	0.05	-0.13	-0.47	0.17	0.53	-0.05	0.52	0.52	0.23	0.47	0.34	0.69	1.0	0.11
total_cases_per_million	0.65	0.58	0.61	0.02	-0.25	-0.53	0.45	0.64	0.03	0.58	0.58	0.27	0.1	0.28	1.0	0.69	-0.23
number_of_waves	0.32	0.08	0.22	-0.26	-0.16	-0.34	0.12	0.29	-0.16	0.3	0.3	0.44	0.27	1.0	0.28	0.34	-0.04
excess_mortality	-0.34	-0.28	-0.16	-0.04	0.09	0.41	-0.27	-0.34	-0.15	-0.27	-0.3	-0.12	1.0	0.27	0.1	0.47	0.55
reproduction_rate	0.5	0.48	0.27	0.05	-0.27	-0.39	0.41	0.5	0.0	0.44	0.45	1.0	-0.12	0.44	0.27	0.23	-0.07
aged_65_older	0.78	0.72	0.62	-0.06	-0.34	-0.57	0.49	0.91	0.06	0.99	1.0	0.45	-0.3	0.3	0.58	0.52	-0.14
aged_70_older	0.76	0.71	0.62	-0.09	-0.35	-0.55	0.47	0.89	0.03	1.0	0.99	0.44	-0.27	0.3	0.58	0.52	-0.13
population_density	0.17	0.23	0.09	0.01	-0.17	-0.03	0.4	0.14	1.0	0.03	0.06	0.0	-0.15	-0.16	0.03	-0.05	-0.08
median_age	0.89	0.84	0.78	0.13	-0.34	-0.69	0.64	1.0	0.14	0.89	0.91	0.5	-0.34	0.29	0.64	0.53	-0.18
gdp_per_capita	0.75	0.68	0.65	0.12	-0.47	-0.5	1.0	0.64	0.4	0.47	0.49	0.41	-0.27	0.12	0.45	0.17	-0.26
extreme_poverty	-0.77	-0.74	-0.75	-0.37	0.18	1.0	-0.5	-0.69	-0.03	-0.55	-0.57	-0.39	0.41	-0.34	-0.53	-0.47	0.0
cardiovasc_death_rate	-0.42	-0.46	0.0	0.14	1.0	0.18	-0.47	-0.34	-0.17	-0.35	-0.34	-0.27	0.09	-0.16	-0.25	-0.13	0.25
diabetes_prevalence	0.19	0.18	0.46	1.0	0.14	-0.37	0.12	0.13	0.01	-0.09	-0.06	0.05	-0.04	-0.26	0.02	0.05	0.0
handwashing_facilities	0.83	0.8	1.0	0.46	0.0	-0.75	0.65	0.78	0.09	0.62	0.62	0.27	-0.16	0.22	0.61	0.56	-0.04
life_expectancy	0.91	1.0	0.8	0.18	-0.46	-0.74	0.68	0.84	0.23	0.71	0.72	0.48	-0.28	0.08	0.58	0.44	-0.17
human_development_index	1.0	0.91	0.83	0.19	-0.42	-0.77	0.75	0.89	0.17	0.76	0.78	0.5	-0.34	0.32	0.65	0.5	-0.24

Heatmaps are a good way to demonstrate a correlation map of a collection of features

12- parallel coordiantes on some features

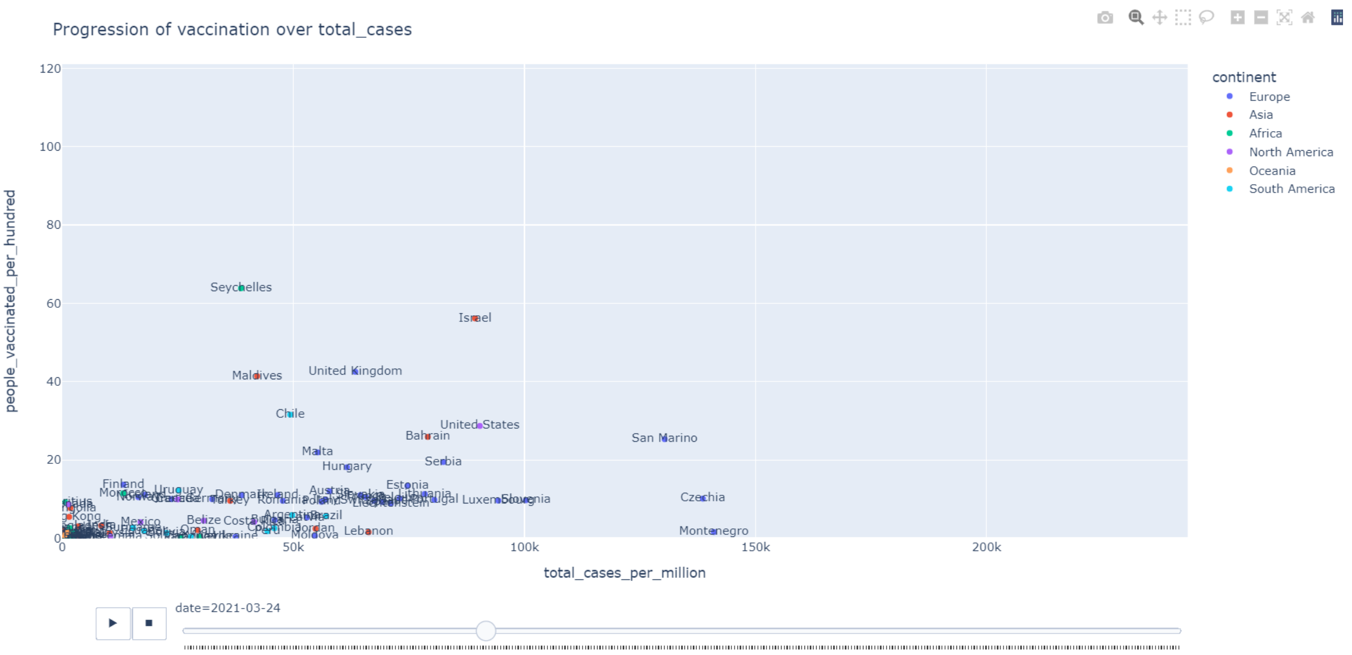
0 : Africa, 1:Asia, 2:Europe, 3:North America, 4:Oceania, 5:South America

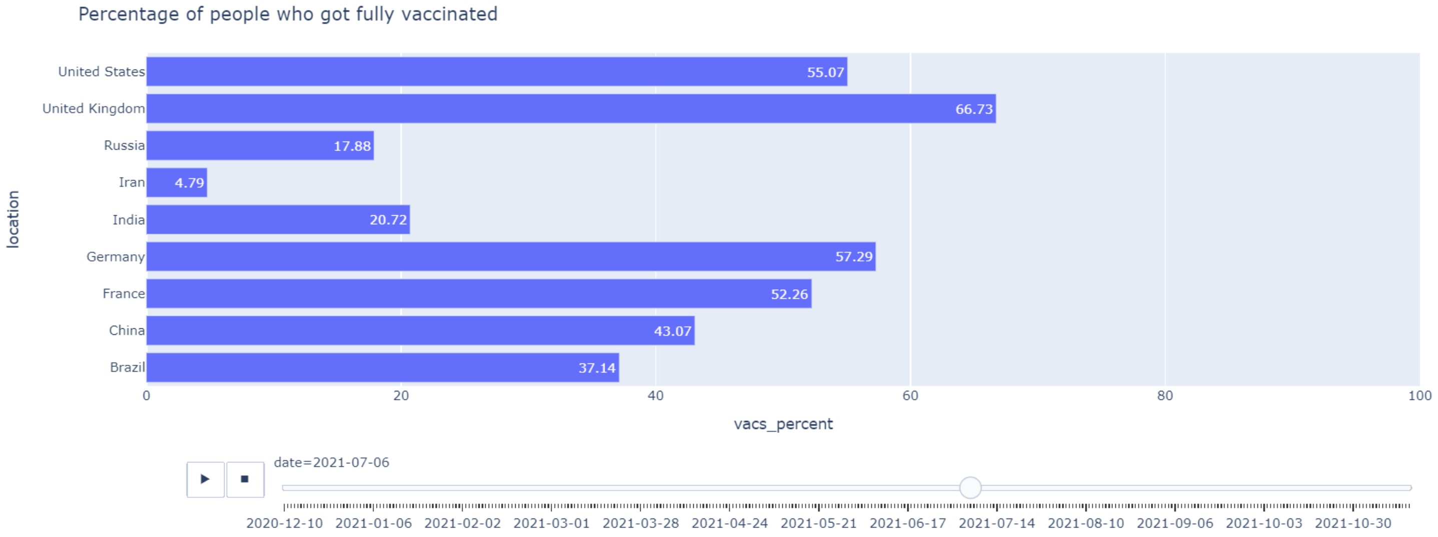


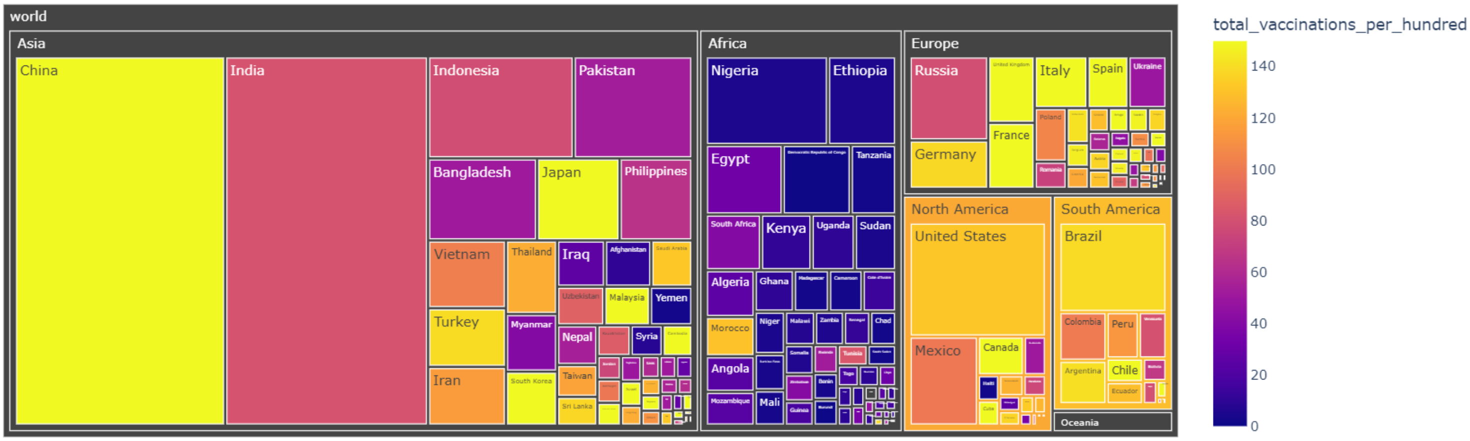
Using a prallel coordinate plot can give us very good relations between features, ofcourse here I couldn't find the best features to use as my plot axis but you get the idea.. the same purpose could be accomplished with a simple scatter plot with connected dots though when we have very different ranges for our features the most dominent one overshadows all the other features, so we use a parallel coordinate system which solves the issue by using a distinct axis for each feature

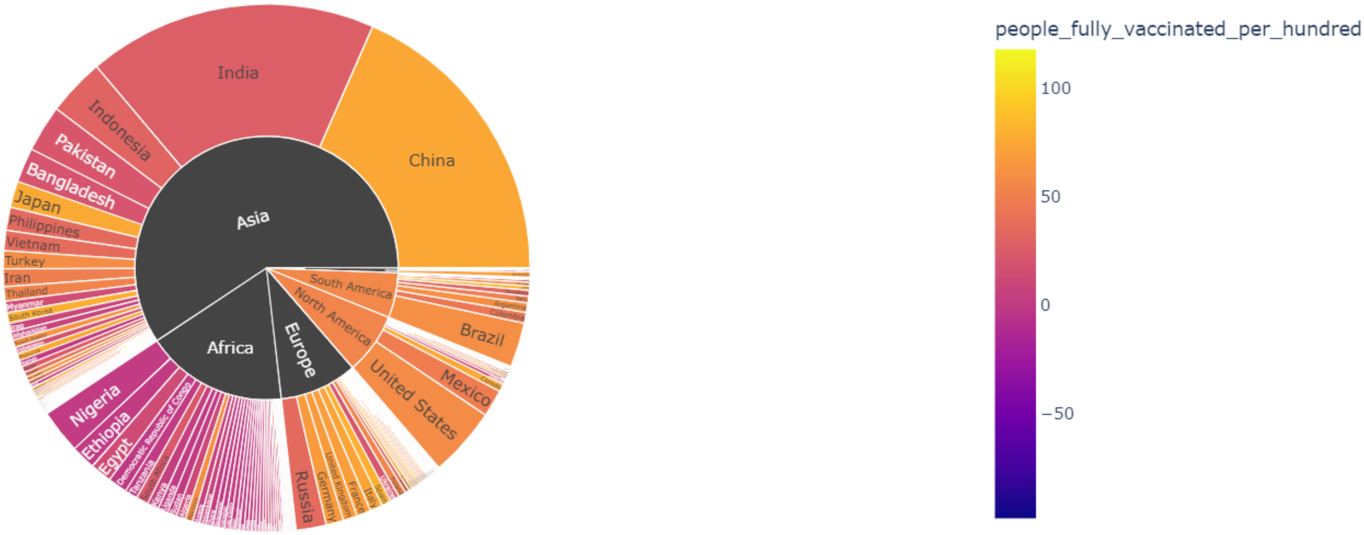
13-vaccinations

We’ll have 4 different visualizaitons to see the changes and effects of vaccination to other features, (other features like hospitalization and icu addmisions and total_tests_per_hundred …. Had too much missing values to be meaningfully plotted)






world
Asia
China
India
Indonesia
Pakistan
Bangladesh
Japan
Philippines
Vietnam
Thailand
Iraq
Afghanistan
Saudi Arabia
Uzbekistan
Malaysia
Yemen
Turkey
Myanmar
Nepal
Syria
Iran
South Korea
Taiwan
Sri Lanka
Africa
Nigeria
Ethiopia
Egypt
Democratic Republic of Congo
Tanzania
South Africa
Kenya
Uganda
Sudan
Algeria
Ghana
Madagascar
Cameroon
Cote d'Ivoire
Morocco
Niger
Malawi
Zambia
Senegal
Chad
Angola
Somalia
Zimbabwe
Tunisia
South Sudan
Benin
Togo
Mozambique
Mali
Guinea
Burundi
Europe
Russia
United Kingdom
Italy
Spain
Ukraine
Poland
France
Germany
Romania
North America
United States
Mexico
Canada
Haiti
Cuba
South America
Brazil
Colombia
Peru
Venezuela
Argentina
Chile
Ecuador
Bolivia
Oceania
total_vaccinations_per_hundred
140
120
100
80
60
40
20
0


India
China
Asia
Indonesia
Pakistan
Bangladesh
Japan
Philippines
Vietnam
Turkey
Iran
Thailand
Myanmar
South Korea
Africa
Nigeria
Ethiopia
Egypt
Democratic Republic of Congo
Tanzania
Kenya
Europe
Russia
Germany
United Kingdom
France
Italy
South America
North America
Brazil
Argentina
Colombia
Mexico
United States
people_fully_vaccinated_per_hundred
100
50
0
−50