Intelligent Data Analys DV1597

Assignment 2

Adnan Altukleh, Abdulkarim Dawalibi

adnantakleh12@gamil.com

karimdawalibi@gmail.com

Background:

Corona

The first case was discovered in China. The new coronavirus was discovered in the city of Wuhan, located in Hubei Province in central China. The infection then spread from Hubei province to other provinces in mainland China and on to a large number of countries in the rest of the world. The disease was declared a pandemic by WHO in mars 2020. Many countries were affected economically, it also changed the way we live. The announcement of the first vaccine was hoped to be a game-changer in the situation and in a matter of short time, everything would get back to normal.

Vaccine

After a vaccination, the body's immune system builds up protection against covid-19. But no vaccine protects one hundred percent. This applies to all vaccines, not just those against covid-19. The main idea of the vaccination is to lower the risk of getting severe symptoms which might lead to death and only getting mild symptoms if the person infected by the virus.

Healthcare

The health service's impact on covid-19 is great. The work of caring for covid patients has meant that resources have had to be adjusted. The capacity of intensive care has increased, and adjustments have been made to new conditions.

Goal of the assignment:

This assignment is about answering questions about covid-19 and the effect of the pandemic on the European countries.

Datasets and tools used:

We were provided three different datasets to extract the answers from during the assignment. The datasets are as follows:

Covid-19: consists of 11 columns and 23649 rows, the most valuable columns contain the reported infected ‘cases’ and the ‘deaths’ of each country from 2020-01 to 2022-04.

Vaccine: consists of 14 columns and 279437 rows, the most valuable columns which will be used during the assignment are ‘first dose’, ‘second dose’,’ dose additional 1’ and ‘Unknown dose’. The data was collected from 2020-w53 to 2022-w17.

Hospital: consists of 7 columns and 26609 rows, the most valuable columns in this dataset are ‘indicator’ and ‘value’. Indicator shows the type of hospitalization if it is ICU or if it is weekly or daily admissions. The value column on the other hand shows the value of the admissions daily/weekly.

Jupyter Notebook was used to visualize and do the analysis.

Data cleaning:

We started to check if any missing values exist in our datasets, and the test showed that only covid-19 contained missing values in the important columns. The dataset is a time series, so we cannot just drop these rows, that's why we decided to handle these values differently. An interpolation function was used to fill the rows with missing values. Interpolation is a method of generating new data points from a discrete set of existing data points, that is, calculating function values that lie between already known values. The function did not eliminate all the missing values, for this reason, we located the rows that has not been taken care of by the function and gave them the mean of the previous five rows.

The next step was to check if any negative values exist in the data sets. The result showed that again only covid-19 contained negative values. We decided to manage this problem by turning the values into positive using the abs () function because it does not make sense to have negative values in both cases and deaths columns.

The existence of duplicate records was also checked in all datasets and the test showed no existence of duplicates.

Finally, we check if any outliers existed in the datasets. The boxplot showed that there are outliers in all columns that were checked. We checked the source and, on the internet, if these outliers(values) should be considered as outliers or something else. What we found and thought from the beginning is that this is real-life data, and we cannot consider these values as outliers. For this reason, we decided to leave it as it is.

Note! All the plots and values mentioned in this session can be found in the Notebook

Q1:

The question was to list the top 10 countries' reported cases in each month of each year.

We started to group the data based on the following columns (month, year, countries and territories and popData2020) and summed up the cases. Then we sorted the data based on year, month and cases. We formulated and sorted the data to make it easier for us to work with it. In this way when we locate each month, we take only the first 10 rows because the data is already sorted from highest to lowest values. After locating the top 10 reported infected cases each month in all the 3 years, we made a function to display the results in an interactive way. Displaying the data in an interactive way makes it easier to understand and smoother to locate only the specific month of the year to show. See appendix (Pic4)

The second part of question 1 was if we find the numbers relatively high compared to the country's population.

In the beginning, we checked if the numbers were relatively high compared to the country's population by dived the sum of cases for each country with the country's population, so we get the percentage value of how many people were infected in every country. Then we sorted the values to see which country had the highest percentage value at the top of the data frame.

We noticed clearly that the values in Denmark, France and Austria had a high value compared to their population, only in Denmark more than 17% were affected by covid-19 only in one month. See appendix (Pic5)

January 2022 had the most reported affected cases in the European countries. So, we located this month in our data frame and wanted to check what is the geographical connection between these countries that reported the most cases in that month. To make it easier for us to see the connection between the countries we plot the data using a Map and each country was given a colour to indicate the number of cases in each country.

The plot shows that the geographical connection between these countries is that it is close to each other, and the reason behind these high records could be because of the Christmas and new year celebrations. During this season people spend their evenings close to each other around the food table. We can also notice that the farther we get from Germany the fewer are the reported affected cases, this might have occurred because of the lack of covid restrictions in Germany. Another reason behind this high number could be because of the coldness, during this period. The coldness helps the pandemic to spreads among people. See appendix (Pic6) and (Pic7)

Note! All the plots and values mentioned in this session can be found in the Notebook

Q2:

The question was about visualizing the total number of cases and deaths in all countries in 2020-2021.

First, we limited the data for 2020 and 2021, then we grouped the data by countries and summed up the number of cases. We did the same thing again, but we summed up the number of deaths. By doing that, we were able to plot the number of cases and deaths each in separate plot for each country on a map. The plot made it clearer, for eyes to see the countries with most and least cases/deaths. See appendix (Pic8) and (Pic9)

Note! All the plots and values mentioned in this session can be found in the Notebook

Q3:

The question was about finding the top 3 used vaccine brands in EU/EEA.

We started with printing all the unique target groups in the data frame. By checking the data set we recognized that there are several countries which do not a target group of 18 > age. The '18 > age 'target group is the sum of all other target groups under 18 (Age0_4, Age10_14, Age15_17). So, we decided to filter the data frame from duplicates, unwanted values, and target groups to get more useful data for this mission. After that, we calculated the total doses sold to all European countries and grouped them by vaccine brand then sorted the data frame to get the top 3 vaccine brands used in EU/EEA. To see the most used vaccine brands, see appendix (pic1).

In the second part of the question, we will show which countries use these brands. In the beginning, we started with group the data by (country, number of weeks, and vaccine brand) and summed up the (First Dose, Second Dose, Unknown Dose, Dose Additional1). From that, we got the data frame for each country, each week and each vaccine brad, and the sum of each type of dose. Then we dropped the empty rows to ensure that only the used vaccines remained and with empty, we mean that the value of First Dose, Second Dose, Unknown Dose and Dose Additional1 is zero in the same row. We Removed the duplicated vaccines in each country and located only the top 3 vaccine brands (COM, MOD, AZ). We counted the number of top 3 vaccine brands used in each country and added it as a column to the data frame. The next step was to group by country and put the name of the top 3 vaccine brands that the country uses in one raw. All countries have used the top 3 vaccines without Liechtenstein, which used only two of them see appendix (pic2).

Note! All the plots and values mentioned in this session can be found in the Notebook

Q4:

The question was about locating the main target group for every vaccine brand of the top 3 used vaccine brands we displayed in the previous question.

We used the same data frame we had in the previous question due this question is based on the last one.

Started by filtering the data frame to ensure that the main target groups, ‘all’ and ‘18 > age’ drop if other target groups exist. If we ignored this filter step, we would be a risk of getting an inaccurate result. After the filter, we had a data frame with each country and the main target of each vaccine brand of the top 3. We decided to make an interactive visualization to make it easier to see each country alone. See appendix (Pic10)

Note! All the plots and values mentioned in this session can be found in the Notebook

Q5:

The question was about finding the most sceptical countries towards the first dos of the covid-19 vaccine.

Our strategy was to sum all the first doses for all the countries in a period between 2021-01 and 2021-06. We chose these dates specifically considering that the availability of the vaccine might differentiate from one country to another, therefore these dates were the best option. After calculating the sum of the first doses in each country during this period, we were able to add a new column to the data frame which contained the population for each country. The population numbers were taken from the covid dataset. Adding the new column made us one step away from our target which is to divide the total number of the first dose for each country by its population and then multiply by 100k. See appendix (pic3)

The second part of the question was to evaluate if that had any impact on the hospitalization level.

Our strategy was to use the hospital dataset to see if getting the first dose helped to reduce the pressure on the ICU or not. So, the most sceptical countries should have the most pressure on their ICU and the opposite on the least sceptical. We started by filtering the dataset to only contain data for the same period we used in the previous part of the question. Then we plotted the daily ICU occupancy of each country divided by the population and multiplied by 100k, to get the daily ICU occupancy per 100k. The plots did not show any remarkable effect on the ICU between the least/most sceptical countries. See appendix (pic11)

These results may have occurred due to several factors, I thought it is good to mention two of them. The first factor is the appearance of new mutated variants of the virus and the vaccine was not so effective. The second factor is that both the new year and Easter holidays increased the contact between people which increased the risk to get infected by the virus. See appendix (pic12)

Note! All the plots and values mentioned in this session can be found in the Notebook

Q6:

The question was about ranking the countries by their vaccinated population under 18 for the first dose of the covid vaccine.

For the first step, we grouped the data based on country, target group, and region and summed the first dose. We chose region because we noticed each country has many regions i.e., Sweden is divided into many regions. But we also noticed if we used the region with the country code, we would get the sum of all the vaccinated in the whole country.

After summing the first dose we took care of the duplicates in target groups under 18 or the All group. Just to make sure we get an accurate result. To get a correct rank we had to divide each sum of the first dose for each country by each country’s population, sorting the result gave us the desired answer.

We plotted the results with a bar plot which showed the most to the least young, vaccinated population for each country. See appendix (pic13)

Note! All the plots and values mentioned in this session can be found in the Notebook

Q7:

The question was about ranking the countries by their vaccinated population over 60 for the second dose of the covid vaccine.

We did pretty much the same as in question 6 only difference is that we ranked the aged, vaccinated population. See appendix (pic14)

Note! All the plots and values mentioned in this session can be found in the Notebook

Q8:

The question was about analysing which countries' health care were most affected in 2020 compared to other countries.

Our strategy was to calculate the total daily hospital admissions divided by the population in each country and visualize the result for the year 2020. We noticed that the data frame is not well structured. We see that daily hospital occupancy is the most common indicator in all the countries. Some countries missed some indicators that is why we chose the most common indicator.

The daily hospital occupancy column showed how many people were in the hospital in a day, so we needed to calculate only the new hospital admissions to output an accurate result.

We Added a population column and divide the total of hospital admissions by population so we get the percentage value which we could compare with another country.

we plotted all countries and the percentage value using a bar plot so it would be easy to locate which country had the worst situation. See appendix (pic15)

Note! All the plots and values mentioned in this session can be found in the Notebook

Discuss:

The results of the answers shows that we are far away from zero covid case in all countries in the nearly future due to the high number of people who still sceptical about the vaccination. The results of some questions cannot be more accurate due to the number of missing values in some columns such as refused doses in vaccine dataset. In Germany they did not have a detailed target group to determine the exact target group for each of the top selling vaccine brands. Due to lack of time, we did not have the chance to answer our questions we thought to write for this analysis. as follows are our questions:

How would the completed vaccination of the population affect the recorded infected cases after 6 months?

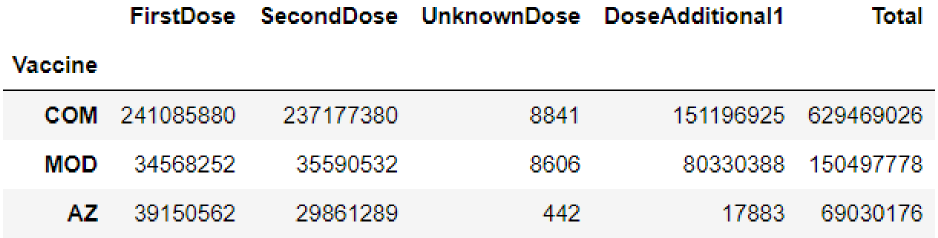
Which vaccine brand sold the most vaccine doses to all EU/EEA, and what is the relationship behind that? Is it the trust factor of the brand that made the brand sell the most?

Being a small country would reduce the number of infected cases compared to the larger countries i.e., Luxembourg and Germany?

Are the eastern European countries more immune than the western countries? What effect could it add to the healthcare system?

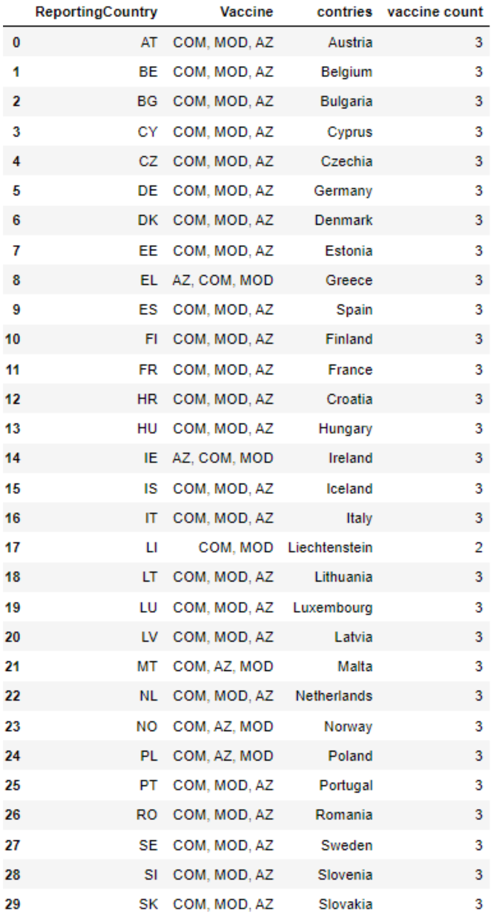
Appendix:

Pic1:



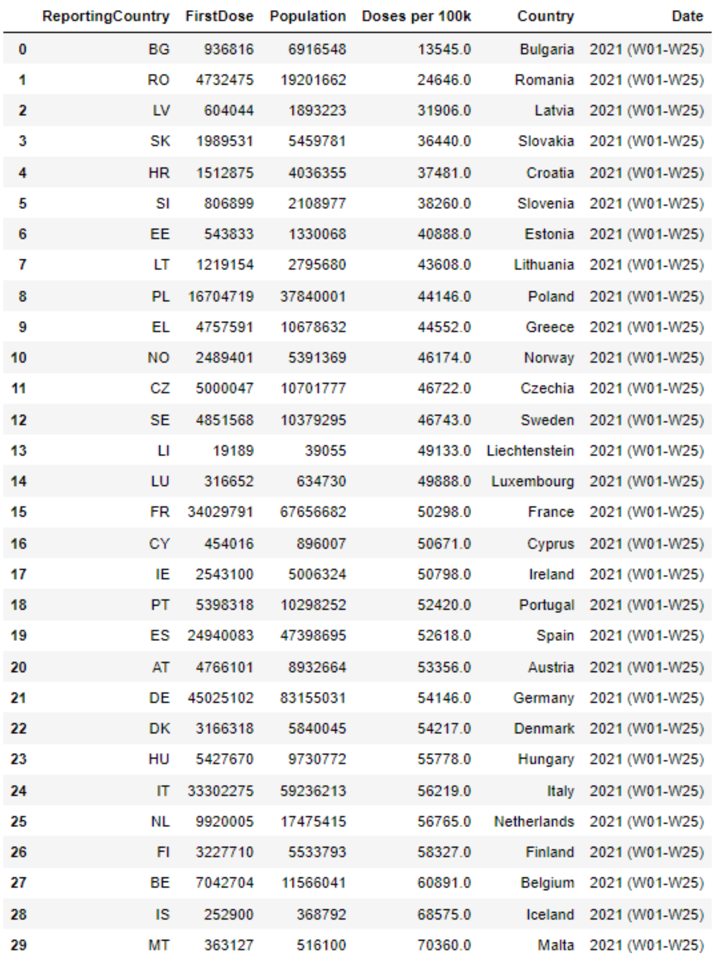
	FirstDose	SecondDose	UnknownDose	DoseAdditional1	Total
Vaccine					
COM	241085880	237177380	8841	151196925	629469026
MOD	34568252	35590532	8606	80330388	150497778
AZ	39150562	29861289	442	17883	69030176

Pic2:



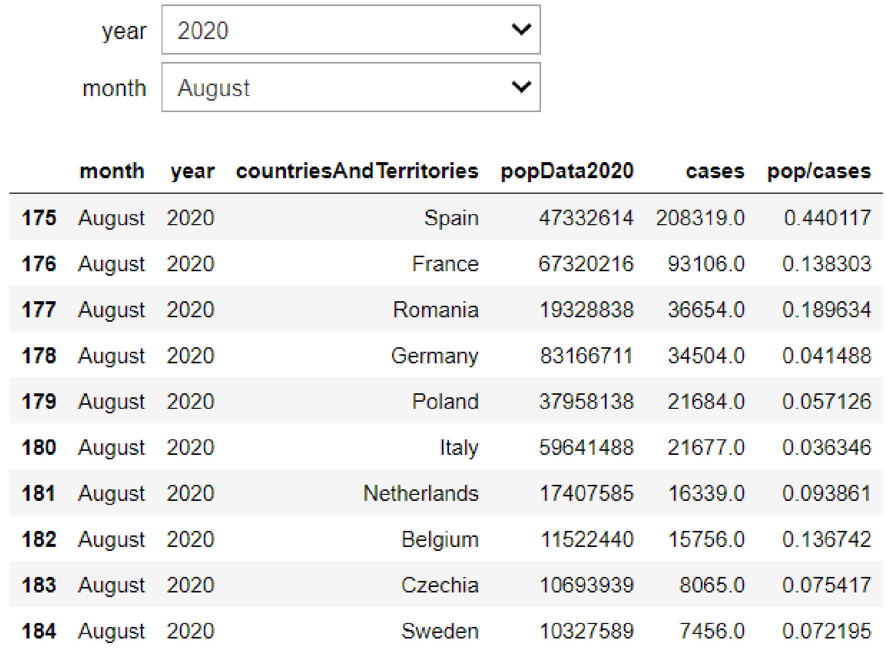
	ReportingCountry	Vaccine	contries	vaccine count
0	AT	COM, MOD, AZ	Austria	3
1	BE	COM, MOD, AZ	Belgium	3
2	BG	COM, MOD, AZ	Bulgaria	3
3	CY	COM, MOD, AZ	Cyprus	3
4	CZ	COM, MOD, AZ	Czechia	3
5	DE	COM, MOD, AZ	Germany	3
6	DK	COM, MOD, AZ	Denmark	3
7	EE	COM, MOD, AZ	Estonia	3
8	EL	AZ, COM, MOD	Greece	3
9	ES	COM, MOD, AZ	Spain	3
10	FI	COM, MOD, AZ	Finland	3
11	FR	COM, MOD, AZ	France	3
12	HR	COM, MOD, AZ	Croatia	3
13	HU	COM, MOD, AZ	Hungary	3
14	IE	AZ, COM, MOD	Ireland	3
15	IS	COM, MOD, AZ	Iceland	3
16	IT	COM, MOD, AZ	Italy	3
17	LI	COM, MOD	Liechtenstein	2
18	LT	COM, MOD, AZ	Lithuania	3
19	LU	COM, MOD, AZ	Luxembourg	3
20	LV	COM, MOD, AZ	Latvia	3
21	MT	COM, AZ, MOD	Malta	3
22	NL	COM, MOD, AZ	Netherlands	3
23	NO	COM, AZ, MOD	Norway	3
24	PL	COM, AZ, MOD	Poland	3
25	PT	COM, MOD, AZ	Portugal	3
26	RO	COM, MOD, AZ	Romania	3
27	SE	COM, MOD, AZ	Sweden	3
28	SI	COM, MOD, AZ	Slovenia	3
29	SK	COM, MOD, AZ	Slovakia	3

Pic3:



	ReportingCountry	FirstDose	Population	Doses per 100k	Country	Date
0	BG	936816	6916548	13545.0	Bulgaria	2021 (W01-W25)
1	RO	4732475	19201662	24646.0	Romania	2021 (W01-W25)
2	LV	604044	1893223	31906.0	Latvia	2021 (W01-W25)
3	SK	1989531	5459781	36440.0	Slovakia	2021 (W01-W25)
4	HR	1512875	4036355	37481.0	Croatia	2021 (W01-W25)
5	SI	806899	2108977	38260.0	Slovenia	2021 (W01-W25)
6	EE	543833	1330068	40888.0	Estonia	2021 (W01-W25)
7	LT	1219154	2795680	43608.0	Lithuania	2021 (W01-W25)
8	PL	16704719	37840001	44146.0	Poland	2021 (W01-W25)
9	EL	4757591	10678632	44552.0	Greece	2021 (W01-W25)
10	NO	2489401	5391369	46174.0	Norway	2021 (W01-W25)
11	CZ	5000047	10701777	46722.0	Czechia	2021 (W01-W25)
12	SE	4851568	10379295	46743.0	Sweden	2021 (W01-W25)
13	LI	19189	39055	49133.0	Liechtenstein	2021 (W01-W25)
14	LU	316652	634730	49888.0	Luxembourg	2021 (W01-W25)
15	FR	34029791	67656682	50298.0	France	2021 (W01-W25)
16	CY	454016	896007	50671.0	Cyprus	2021 (W01-W25)
17	IE	2543100	5006324	50798.0	Ireland	2021 (W01-W25)
18	PT	5398318	10298252	52420.0	Portugal	2021 (W01-W25)
19	ES	24940083	47398695	52618.0	Spain	2021 (W01-W25)
20	AT	4766101	8932664	53356.0	Austria	2021 (W01-W25)
21	DE	45025102	83155031	54146.0	Germany	2021 (W01-W25)
22	DK	3166318	5840045	54217.0	Denmark	2021 (W01-W25)
23	HU	5427670	9730772	55778.0	Hungary	2021 (W01-W25)
24	IT	33302275	59236213	56219.0	Italy	2021 (W01-W25)
25	NL	9920005	17475415	56765.0	Netherlands	2021 (W01-W25)
26	FI	3227710	5533793	58327.0	Finland	2021 (W01-W25)
27	BE	7042704	11566041	60891.0	Belgium	2021 (W01-W25)
28	IS	252900	368792	68575.0	Iceland	2021 (W01-W25)
29	MT	363127	516100	70360.0	Malta	2021 (W01-W25)

Pic4:

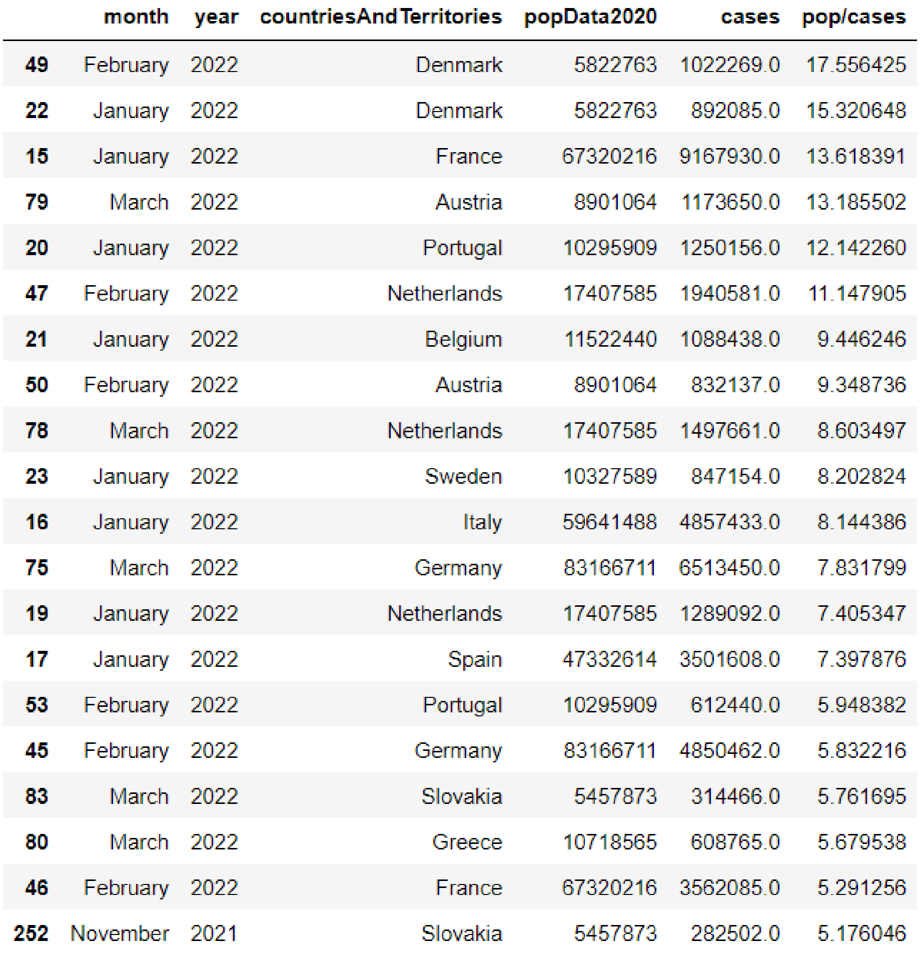


year 2020

month August

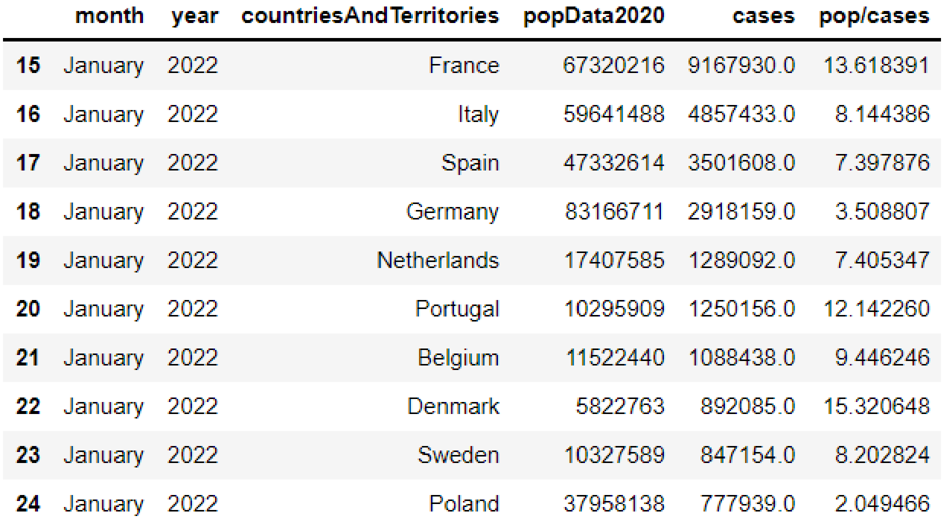
	month	year	countriesAndTerritories	popData2020	cases	pop/cases
175	August	2020	Spain	47332614	208319.0	0.440117
176	August	2020	France	67320216	93106.0	0.138303
177	August	2020	Romania	19328838	36654.0	0.189634
178	August	2020	Germany	83166711	34504.0	0.041488
179	August	2020	Poland	37958138	21684.0	0.057126
180	August	2020	Italy	59641488	21677.0	0.036346
181	August	2020	Netherlands	17407585	16339.0	0.093861
182	August	2020	Belgium	11522440	15756.0	0.136742
183	August	2020	Czechia	10693939	8065.0	0.075417
184	August	2020	Sweden	10327589	7456.0	0.072195

Pic5:



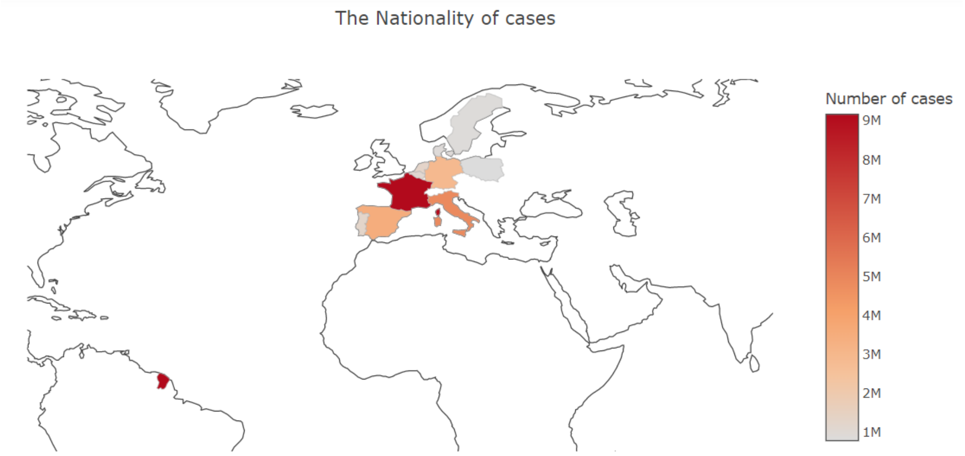
	month	year	countriesAndTerritories	popData2020	cases	pop/cases
49	February	2022	Denmark	5822763	1022269.0	17.556425
22	January	2022	Denmark	5822763	892085.0	15.320648
15	January	2022	France	67320216	9167930.0	13.618391
79	March	2022	Austria	8901064	1173650.0	13.185502
20	January	2022	Portugal	10295909	1250156.0	12.142260
47	February	2022	Netherlands	17407585	1940581.0	11.147905
21	January	2022	Belgium	11522440	1088438.0	9.446246
50	February	2022	Austria	8901064	832137.0	9.348736
78	March	2022	Netherlands	17407585	1497661.0	8.603497
23	January	2022	Sweden	10327589	847154.0	8.202824
16	January	2022	Italy	59641488	4857433.0	8.144386
75	March	2022	Germany	83166711	6513450.0	7.831799
19	January	2022	Netherlands	17407585	1289092.0	7.405347
17	January	2022	Spain	47332614	3501608.0	7.397876
53	February	2022	Portugal	10295909	612440.0	5.948382
45	February	2022	Germany	83166711	4850462.0	5.832216
83	March	2022	Slovakia	5457873	314466.0	5.761695
80	March	2022	Greece	10718565	608765.0	5.679538
46	February	2022	France	67320216	3562085.0	5.291256
252	November	2021	Slovakia	5457873	282502.0	5.176046

Pic6:

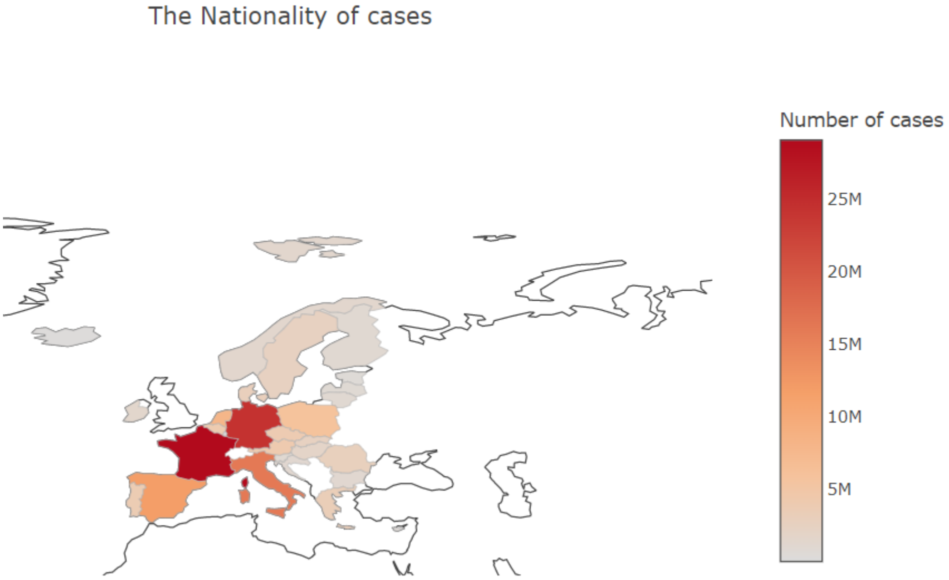


	month	year	countriesAndTerritories	popData2020	cases	pop/cases
15	January	2022	France	67320216	9167930.0	13.618391
16	January	2022	Italy	59641488	4857433.0	8.144386
17	January	2022	Spain	47332614	3501608.0	7.397876
18	January	2022	Germany	83166711	2918159.0	3.508807
19	January	2022	Netherlands	17407585	1289092.0	7.405347
20	January	2022	Portugal	10295909	1250156.0	12.142260
21	January	2022	Belgium	11522440	1088438.0	9.446246
22	January	2022	Denmark	5822763	892085.0	15.320648
23	January	2022	Sweden	10327589	847154.0	8.202824
24	January	2022	Poland	37958138	777939.0	2.049466

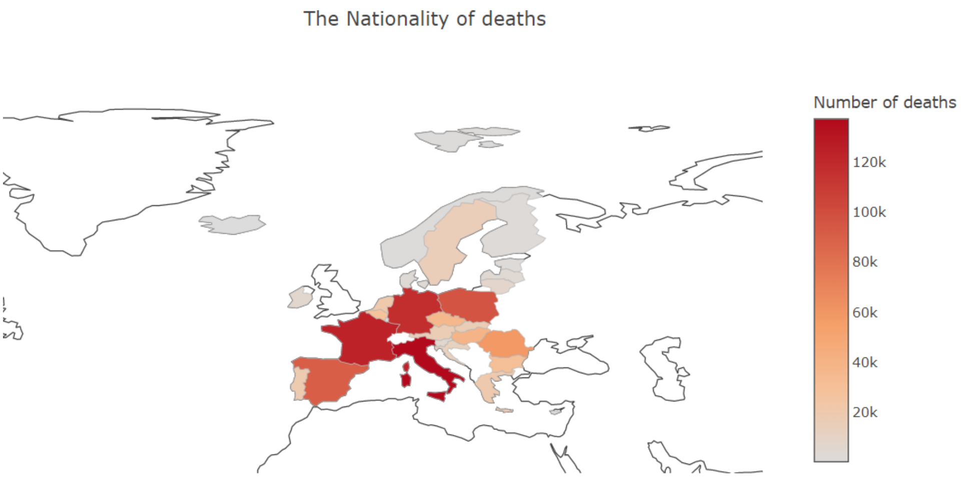
Pic7:


The Nationality of cases
Number of cases
9M
8M
7M
6M
5M
4M
3M
2M
1M

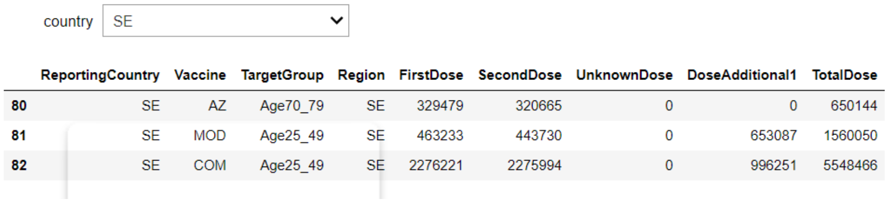
Pic8:


The Nationality of cases
Number of cases
25M
20M
15M
10M
5M

Pic9:



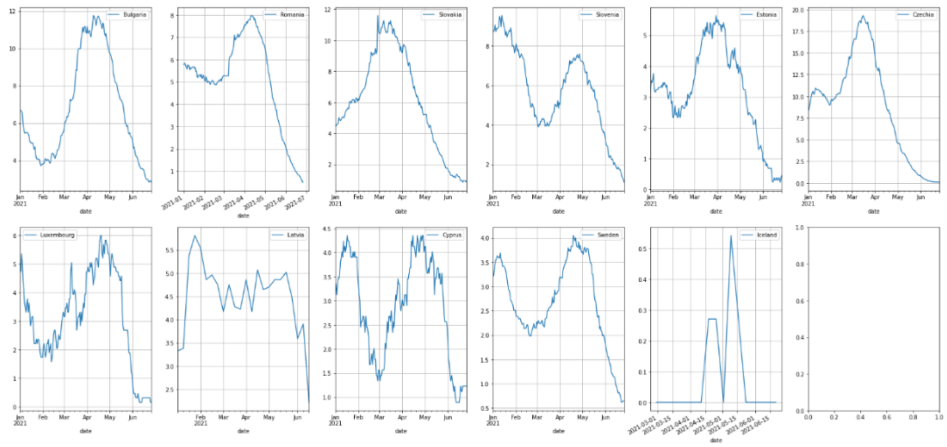
Pic10:



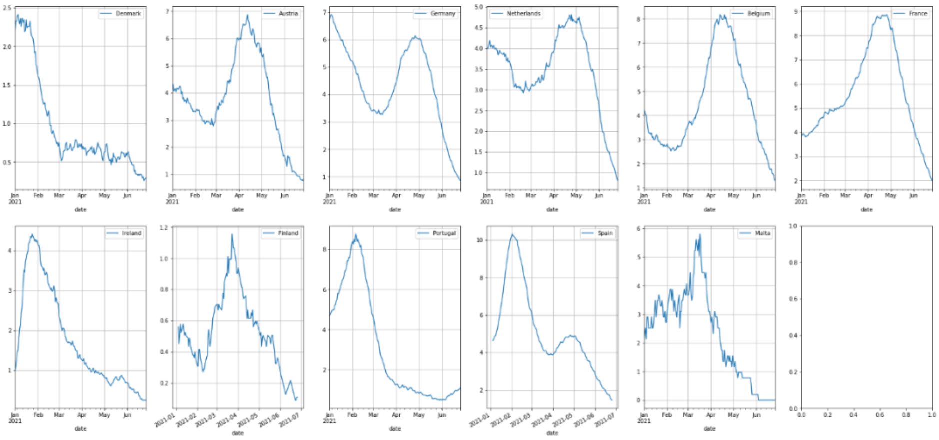
country SE

	ReportingCountry	Vaccine	TargetGroup	Region	FirstDose	SecondDose	UnknownDose	DoseAdditional1	TotalDose
80	SE	AZ	Age70_79	SE	329479	320665	0	0	650144
81	SE	MOD	Age25_49	SE	463233	443730	0	653087	1560050
82	SE	COM	Age25_49	SE	2276221	2275994	0	996251	5548466

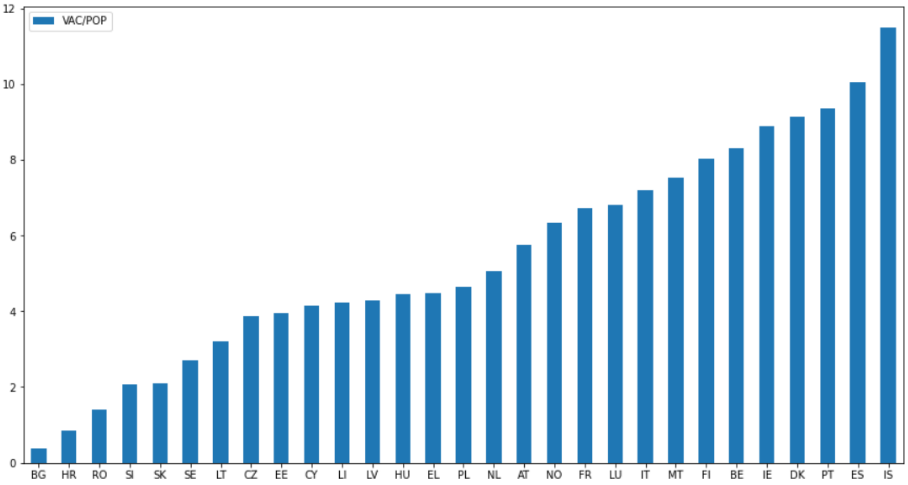
Pic11:



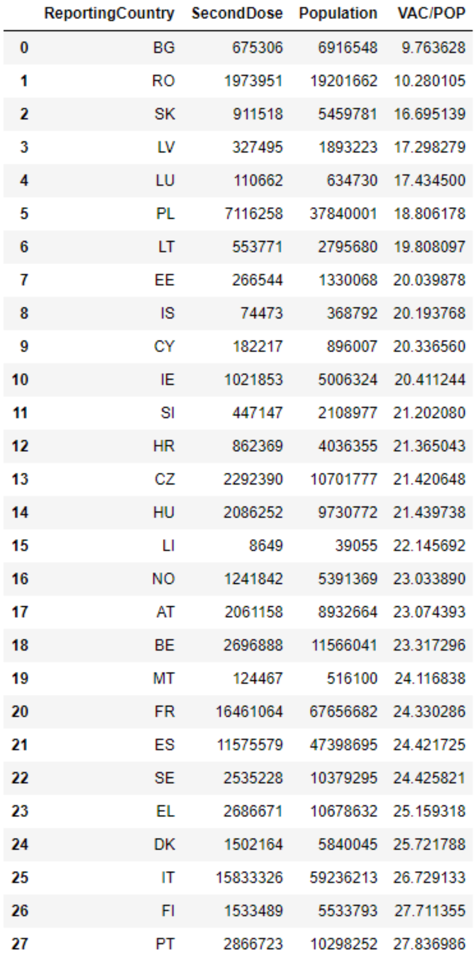
Pic12:


Denmark
Austria
Germany
Netherlands
Belgium
France
Ireland
Finland
Portugal
Spain
Malta
date

Pic13:

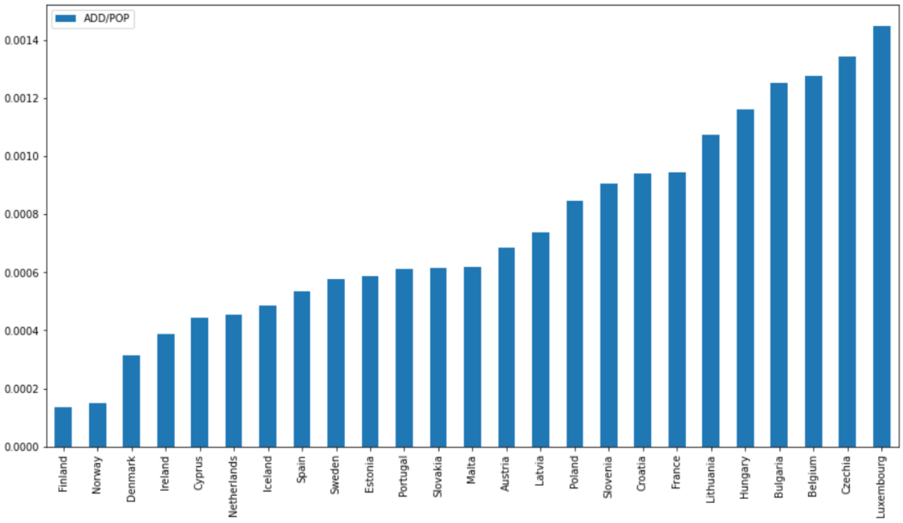

VAC/POP
0
2
4
6
8
10
12
BG
HR
RO
SI
SK
SE
LT
CZ
EE
CY
LI
LV
HU
EL
PL
NL
AT
NO
FR
LU
IT
MT
FI
BE
IE
DK
PT
ES
IS

Pic14:



	ReportingCountry	SecondDose	Population	VAC/POP
0	BG	675306	6916548	9.763628
1	RO	1973951	19201662	10.280105
2	SK	911518	5459781	16.695139
3	LV	327495	1893223	17.298279
4	LU	110662	634730	17.434500
5	PL	7116258	37840001	18.806178
6	LT	553771	2795680	19.808097
7	EE	266544	1330068	20.039878
8	IS	74473	368792	20.193768
9	CY	182217	896007	20.336560
10	IE	1021853	5006324	20.411244
11	SI	447147	2108977	21.202080
12	HR	862369	4036355	21.365043
13	CZ	2292390	10701777	21.420648
14	HU	2086252	9730772	21.439738
15	LI	8649	39055	22.145692
16	NO	1241842	5391369	23.033890
17	AT	2061158	8932664	23.074393
18	BE	2696888	11566041	23.317296
19	MT	124467	516100	24.116838
20	FR	16461064	67656682	24.330286
21	ES	11575579	47398695	24.421725
22	SE	2535228	10379295	24.425821
23	EL	2686671	10678632	25.159318
24	DK	1502164	5840045	25.721788
25	IT	15833326	59236213	26.729133
26	FI	1533489	5533793	27.711355
27	PT	2866723	10298252	27.836986

Pic15:


ADD/POP
Finland
Norway
Denmark
Ireland
Cyprus
Netherlands
Iceland
Spain
Sweden
Estonia
Portugal
Slovakia
Malta
Austria
Latvia
Poland
Slovenia
Croatia
France
Lithuania
Hungary
Bulgaria
Belgium
Czechia
Luxembourg