Group Delivery

18/01/2021 to 25/01/2021

Group D

Javier Gil, M. Ángeles Rodríguez & Ariadna Puigventós

Data Science

The Bridge

General Vision

First of all, we couldn't imagine how it was to work with a team because the planning and managing have been a lot of meticulously. We have been dealing a lot of handicaps during the process. Secondly, each one knew their knowledges about code methods and we planned to be efficient and each question wasn't repeat in the document.

Finally, the first day we were analyzing and asking ourself about what the meaning were there in .csv and how it was treaty.

Goals

We set out to reach some questions of Option A goal.

Specifications

Software

- Code Visual Studio
- Google Search
- Git Hub
- Python Tutor
- Pages (another version word)
- Power Point

Hardware

- PC Windows
- MacBook Pro 8GB

Requirements

Our project is uploaded in **GitHub repository** https://github.com/JGILANTU/GroupD_Covid19

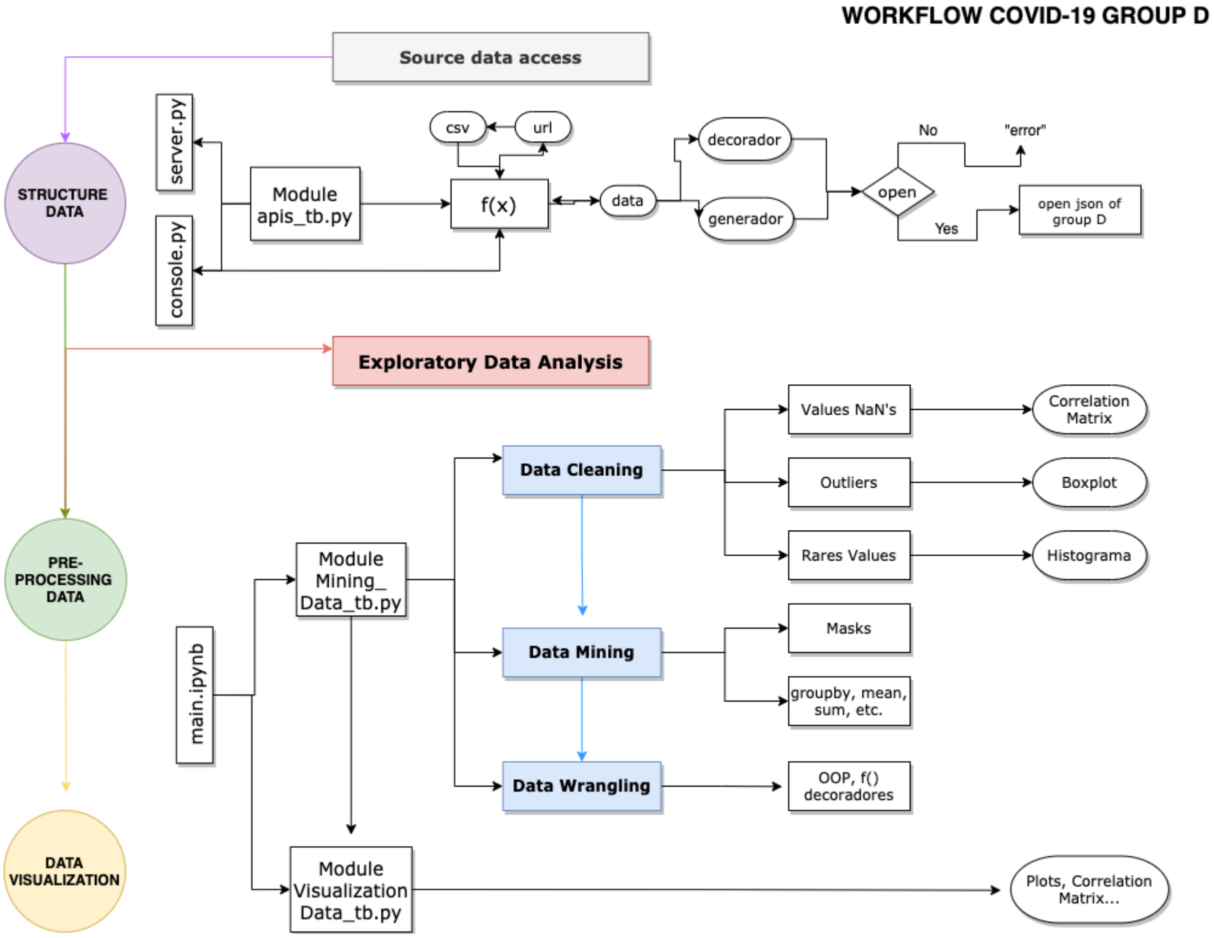
Before starting to project we organized the structure folders and files.

Set Up Structure:

1. We worked with different tests files in notebook/ folder where there are different operations, correct and fail code.

2. When the code returned correctly, it was transferred and copied in each module appropriate with its function in utils/ folders.
3. Imported every function to main.ipynb file in src/ folder.

We created a workflow and Optimization planning that we have been changing during the process:



Peu de foto

Steps

I. Research the context

We was researching about all of state of alarm about our countries: Venezuela, Portugal, Turkia, United Kingdom and Spain.

Then we wanted to research about Venezuela because is one of country with less new cases and new deaths, and sometimes we didn't believe it, but the numbers are numbers :)

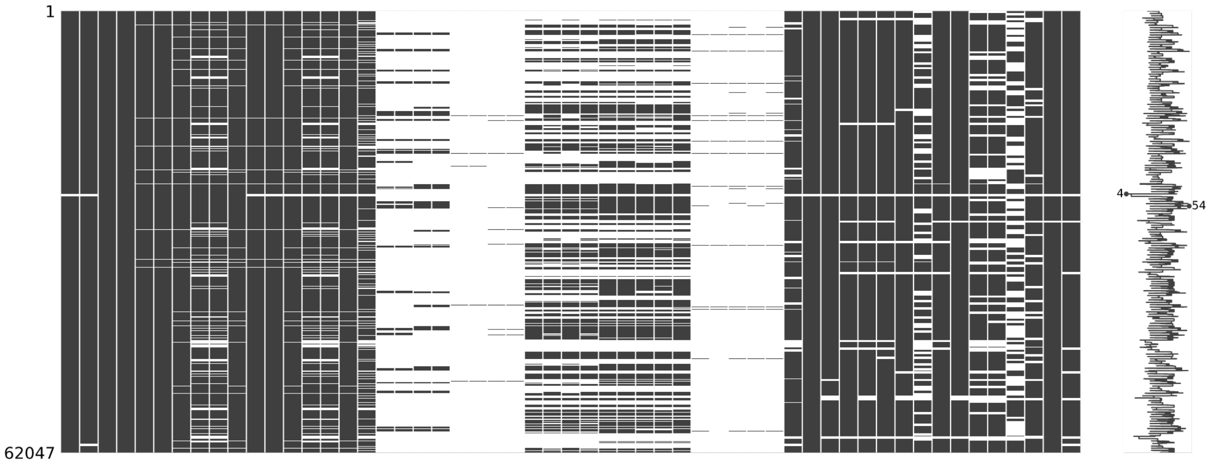
II. Get Data

We get data from the url https://ourworldindata.org/coronavirus-source-data

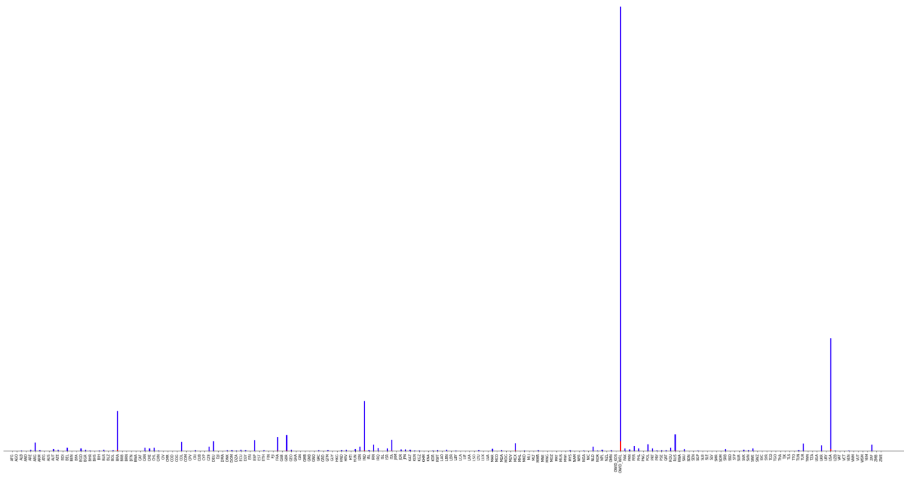
In a particular case, we had to check some data in url https://www.who.int/gho/database/es/ because there was a question about total recoveries but it was impossible to extract this information about the first link, because it was not exist. Then we declined this information for another one.

III. Data Cleaning & Wrangling

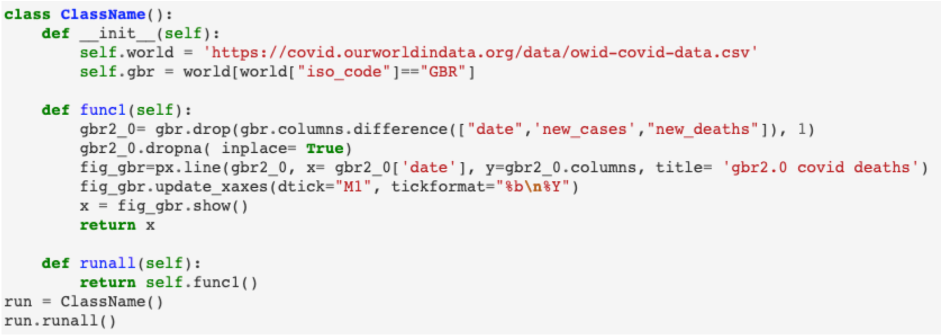
First of all, we wante to see how many NaN's we were talking in our dataframe, then we wanted to show of a visual manner which columns were innecessary because one of the questions was showing different tendencies for each columns.



Secondly, we wanted to detect some outliers or rare values to disrupt or pervert our origin values, we used a boxplot and histogram about the column. Finally, we detected there was a row about total world and this value was not a country:



We created some functions and one OOP to extract one of the tables about one country to see which information there was:

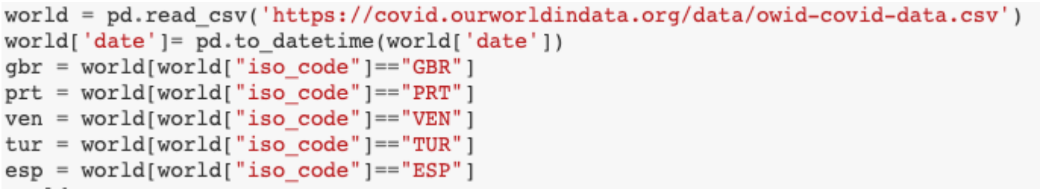


```
class ClassName():
    def __init__(self):
        self.world = 'https://covid.ourworldindata.org/data/owid-covid-data.csv'
        self.gbr = world[world["iso_code"]=="GBR"]

    def func1(self):
        gbr2_0= gbr.drop(gbr.columns.difference(["date",'new_cases',"new_deaths"]), 1)
        gbr2_0.dropna( inplace= True)
        fig_gbr=px.line(gbr2_0, x= gbr2_0['date'], y=gbr2_0.columns, title= 'gbr2.0 covid deaths')
        fig_gbr.update_xaxes(dtick="M1", tickformat="%b\n%Y")
        x = fig_gbr.show()
        return x

    def runall(self):
        return self.func1()
run = ClassName()
run.runall()
```

Furthermore, we had to change some columns from object to datatime or float64 to integer.

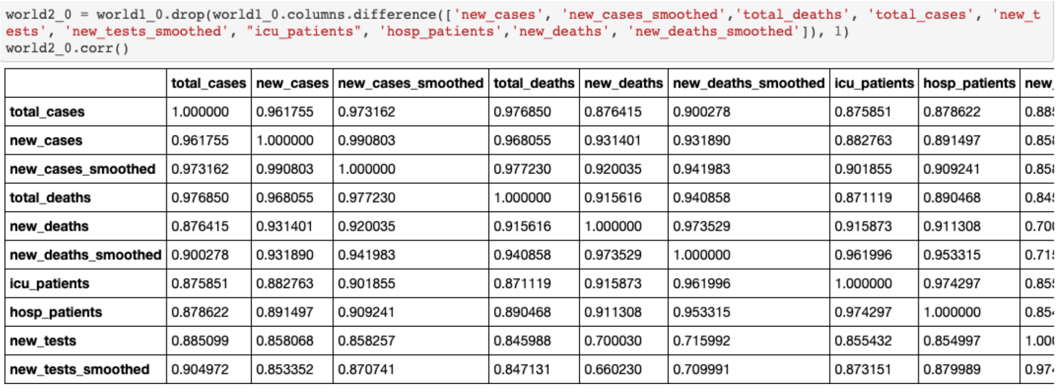
IV. Data Mining

One of example is when we had to make a different mask over origin dataframe to extract only our countries:



```
world = pd.read_csv('https://covid.ourworldindata.org/data/owid-covid-data.csv')
world['date']= pd.to_datetime(world['date'])
gbr = world[world["iso_code"]=="GBR"]
prt = world[world["iso_code"]=="PRT"]
ven = world[world["iso_code"]=="VEN"]
tur = world[world["iso_code"]=="TUR"]
esp = world[world["iso_code"]=="ESP"]
```

Which variables had more correlation positve between them:

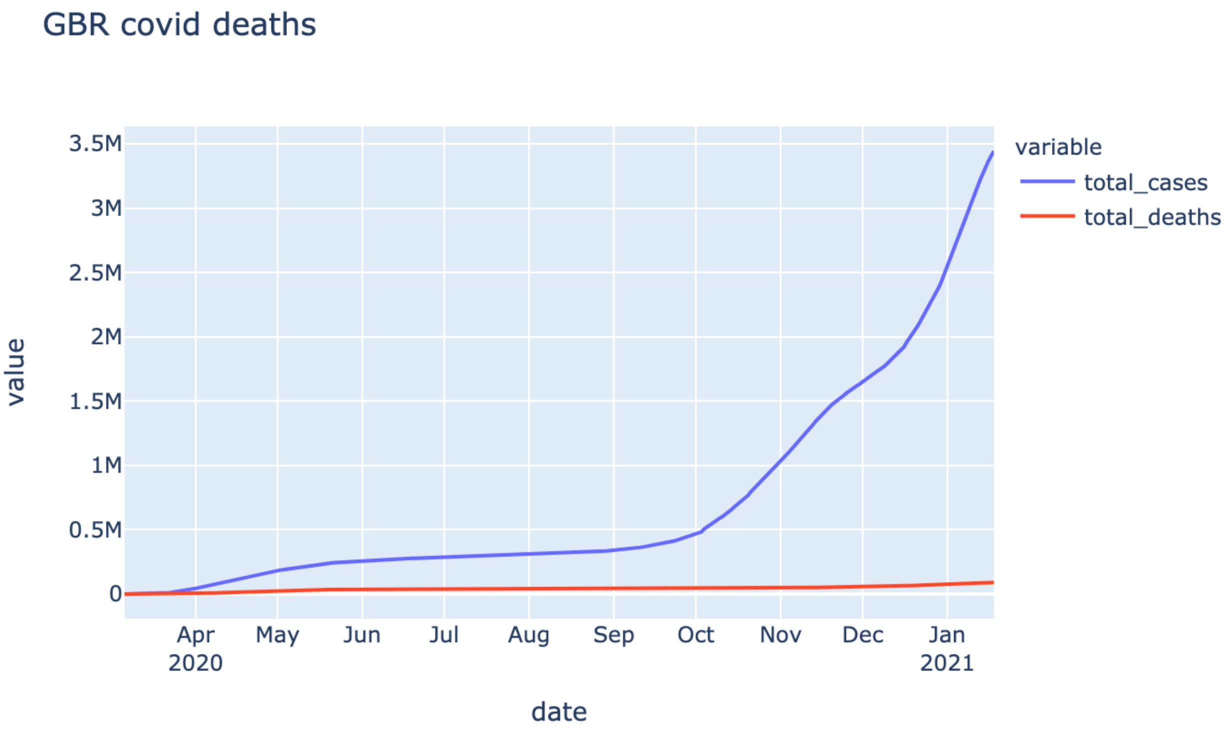


```
world2_0 = world1_0.drop(world1_0.columns.difference(['new_cases', 'new_cases_smoothed','total_deaths', 'total_cases', 'new_t
ests', 'new_tests_smoothed', "icu_patients", 'hosp_patients','new_deaths', 'new_deaths_smoothed']), 1)
world2_0.corr()
```

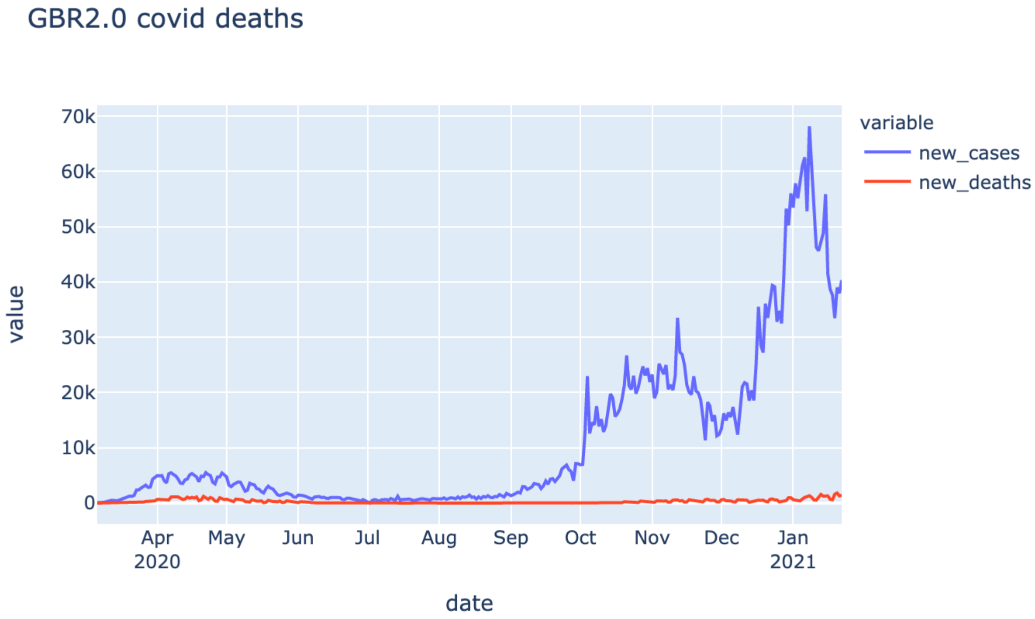
	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	icu_patients	hosp_patients	new_
total_cases	1.000000	0.961755	0.973162	0.976850	0.876415	0.900278	0.875851	0.878622	0.88
new_cases	0.961755	1.000000	0.990803	0.968055	0.931401	0.931890	0.882763	0.891497	0.85
new_cases_smoothed	0.973162	0.990803	1.000000	0.977230	0.920035	0.941983	0.901855	0.909241	0.85
total_deaths	0.976850	0.968055	0.977230	1.000000	0.915616	0.940858	0.871119	0.890468	0.84
new_deaths	0.876415	0.931401	0.920035	0.915616	1.000000	0.973529	0.915873	0.911308	0.70
new_deaths_smoothed	0.900278	0.931890	0.941983	0.940858	0.973529	1.000000	0.961996	0.953315	0.71
icu_patients	0.875851	0.882763	0.901855	0.871119	0.915873	0.961996	1.000000	0.974297	0.85
hosp_patients	0.878622	0.891497	0.909241	0.890468	0.911308	0.953315	0.974297	1.000000	0.85
new_tests	0.885099	0.858068	0.858257	0.845988	0.700030	0.715992	0.855432	0.854997	1.00
new_tests_smoothed	0.904972	0.853352	0.870741	0.847131	0.660230	0.709991	0.873151	0.879989	0.97

V. Data Vizualition

It is showing how total cases increases meanwhile total deaths as well (not parallel).



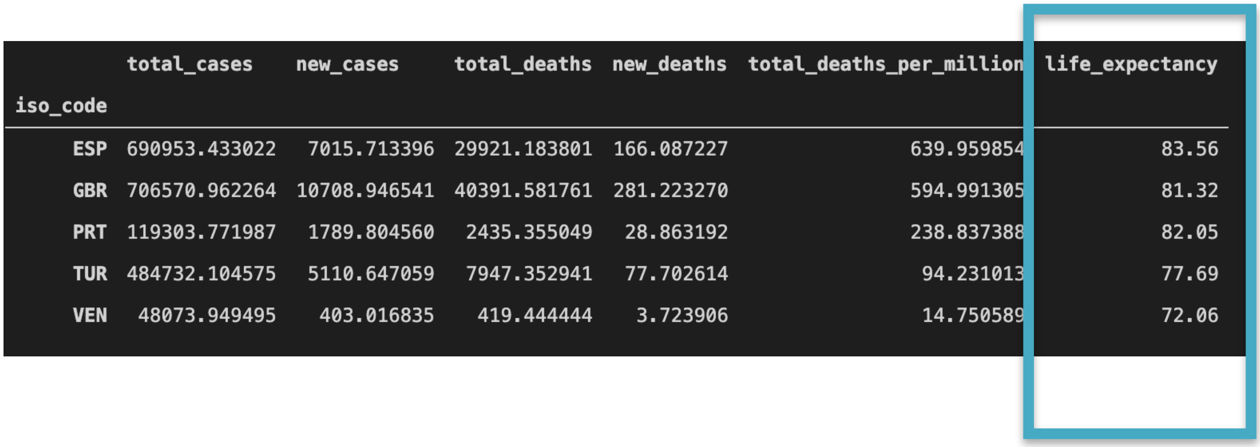
In this case, it's an exemple with how we wanted to get other variable new cases to show how the cases increases every day.



VI. Conclusions

C10) First of all,
it is showing the position of our countries: Portugal, Venezuela, Turkey, UK and Spain in general ranking in over the world. Venezuela is the first country with less new cases and new deaths being 116 position in new cases and 106 in new deaths respect 190 countrie s total. Instead of the last country is United Kingdom in 186 position in new cases and de aths. Secondly,
Spain is the second last country being 182/190 and 180/190 in new cases and deaths re spectively. In inspite of, Spain is the first country (our countries) where the life expectancy is highest with 83.5 years old of mean, the next one is Portugal with 82 year old of mean and the last one is Venezuela with 72 year old of mean.

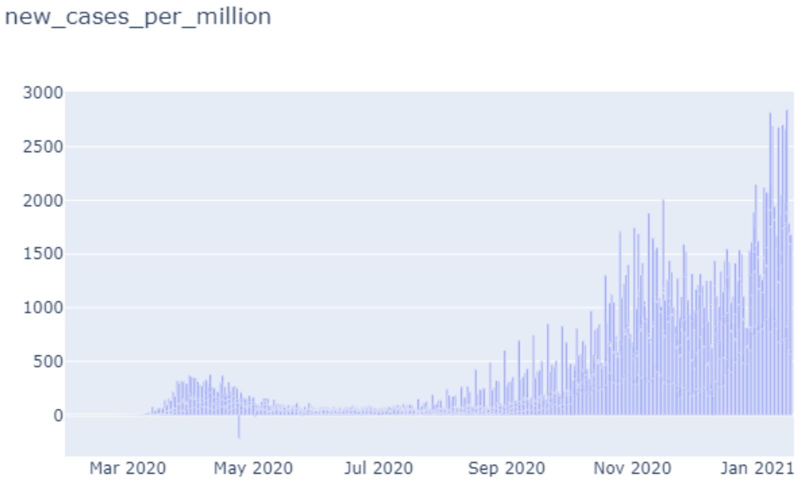
The conclusion is there are not any correlation positive with new cases or new deaths with life expectancy. We would like to know in the future if this life expectancy in each country is better or worst.



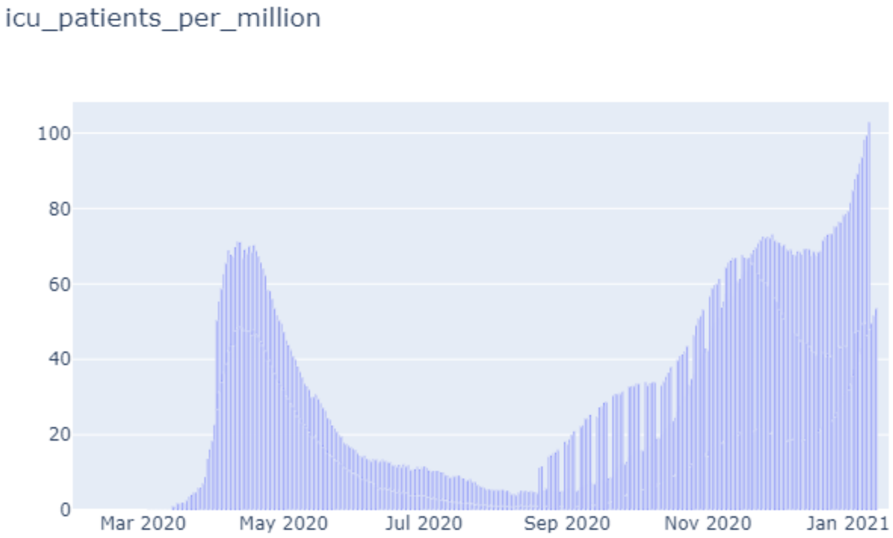
	total_cases	new_cases	total_deaths	new_deaths	total_deaths_per_million	life_expectancy
iso_code						
ESP	690953.433022	7015.713396	29921.183801	166.087227	639.959854	83.56
GBR	706570.962264	10708.946541	40391.581761	281.223270	594.991305	81.32
PRT	119303.771987	1789.804560	2435.355049	28.863192	238.837388	82.05
TUR	484732.104575	5110.647059	7947.352941	77.702614	94.231013	77.69
VEN	48073.949495	403.016835	419.444444	3.723906	14.750589	72.06

B4)

We concluded that for example in our countries the worst moments to go it would be at the begin of summer because there are few cases, but not deaths.



Furthermore, if we can detect by icu patients variable we wouldn’t recommend to travel any period because the hospitals couldn’t attend you is due of hospitals and icu areas would be very crowed.



A4) The state of alarm in our countries didn’t worked pretty much because the variables about new cases and new deaths was increasing during pandemic, either way we have not had much information to concluded 100%.