Covid-19 in Europe: analysis of patterns and predictions for the coming weeks.

Autor: *Ing. Miroshnitshenko Patricio*

Summary:

In the world, the covid is on the increase, although in regions there were decreases in cases, we find continents such as Europe where a second wave of infections is being seen due to the decrease in restrictions, the increase in tourism and regional movement.

Knowing the current problem that is being experienced in all continents: this report was made to show where the greatest increases in covid-19 cases are occurring, with the aim of recognizing some patterns that allow us to visualize how the current situation and predict short-term

situations,

- This study was based primarily on data from:
- johns hopskins University https://coronavirus.jhu.edu/map.html
- World Health Organization https://www.who.int/. World Data bank
- https://data.worldbank.org/

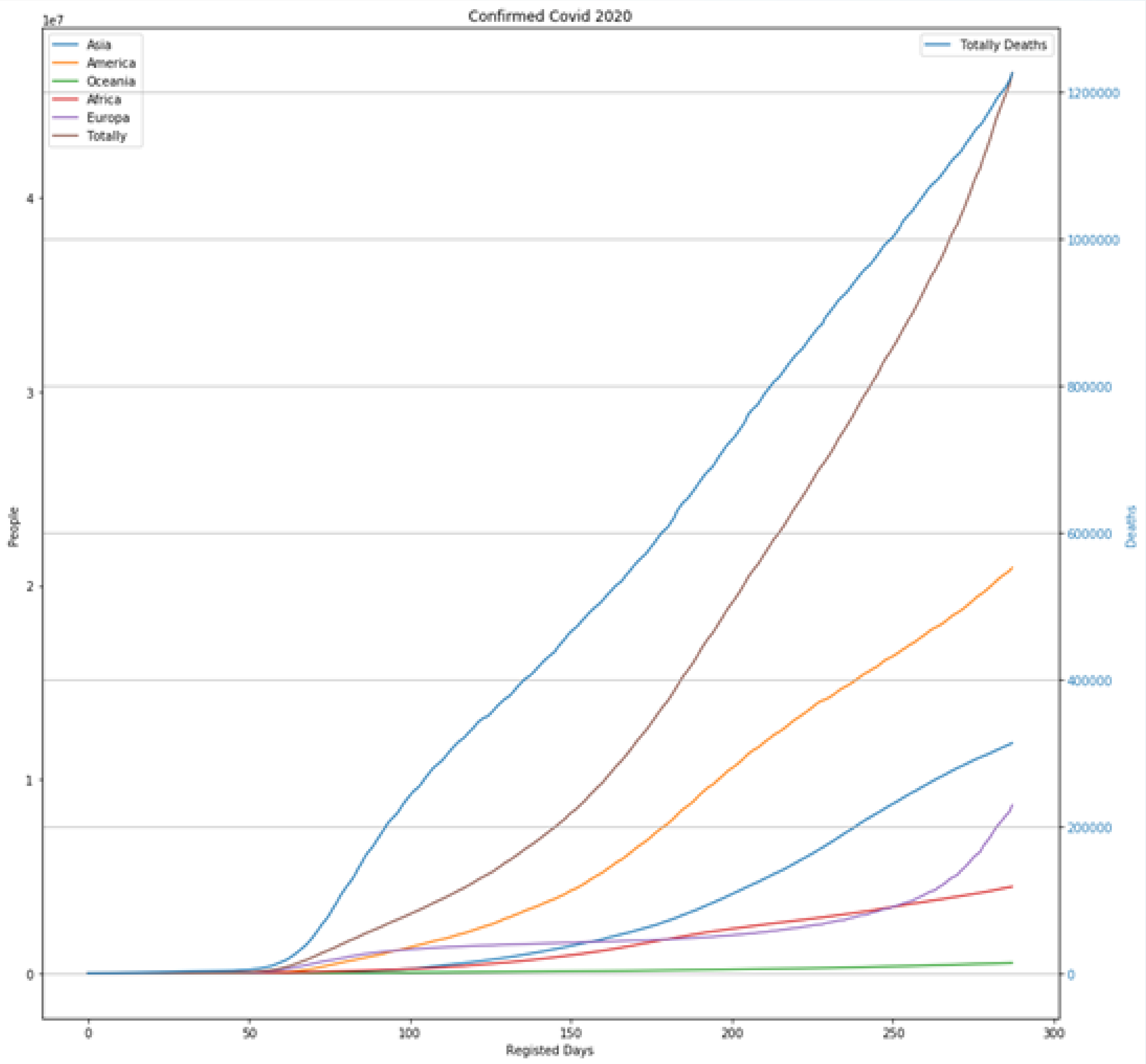
The data collection, preparation, cleaning, treatment is in the attached python file where it is shown and explains in part how the data was treated for use in this report. In addition to all the graphics used in this Report.

This report was made with the data up to the date of November 4, 2020 and refers to that day as the current one.

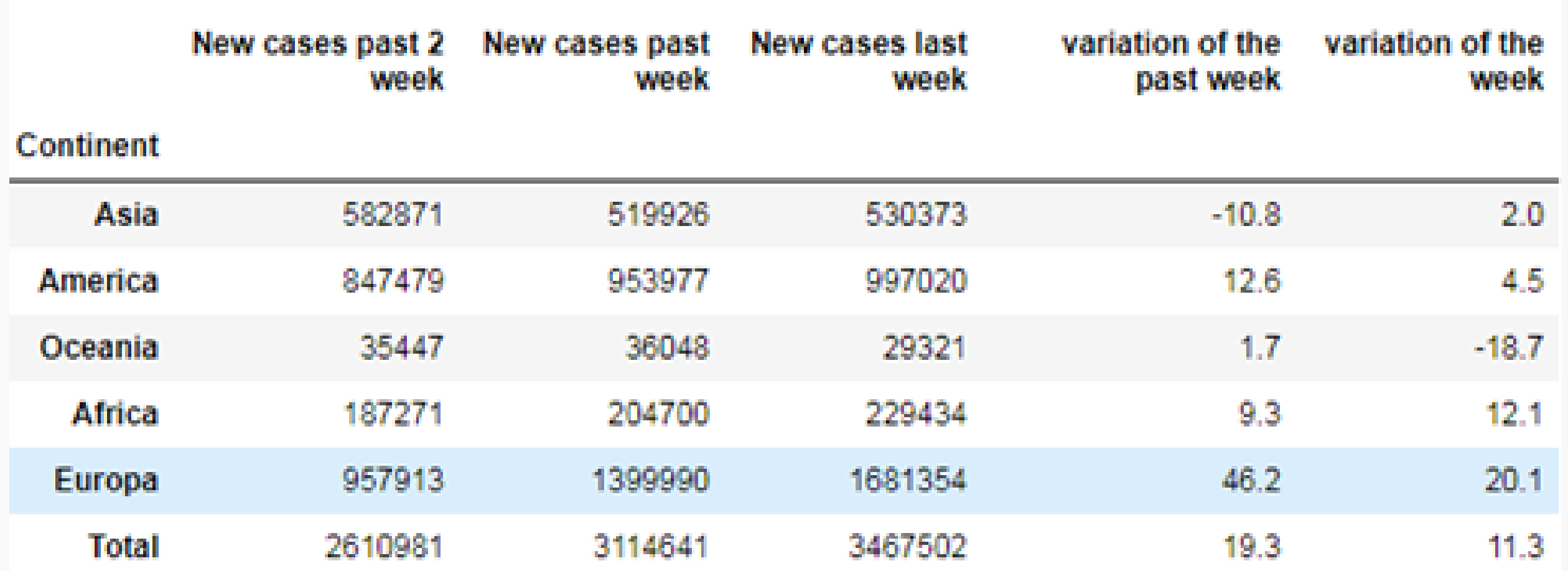
Introduction:

the levels of infections in the world are currently growing exponentially, which leads to more and more confirmed cases in recent weeks. This translates to a greater number of deaths in the next few weeks,

The objective of governments at present is to stabilize this contagion curve, in order not to saturate the health system and have advantages in relation to the time until a vaccine is found. Currently in the world the cases by continent are as follows:

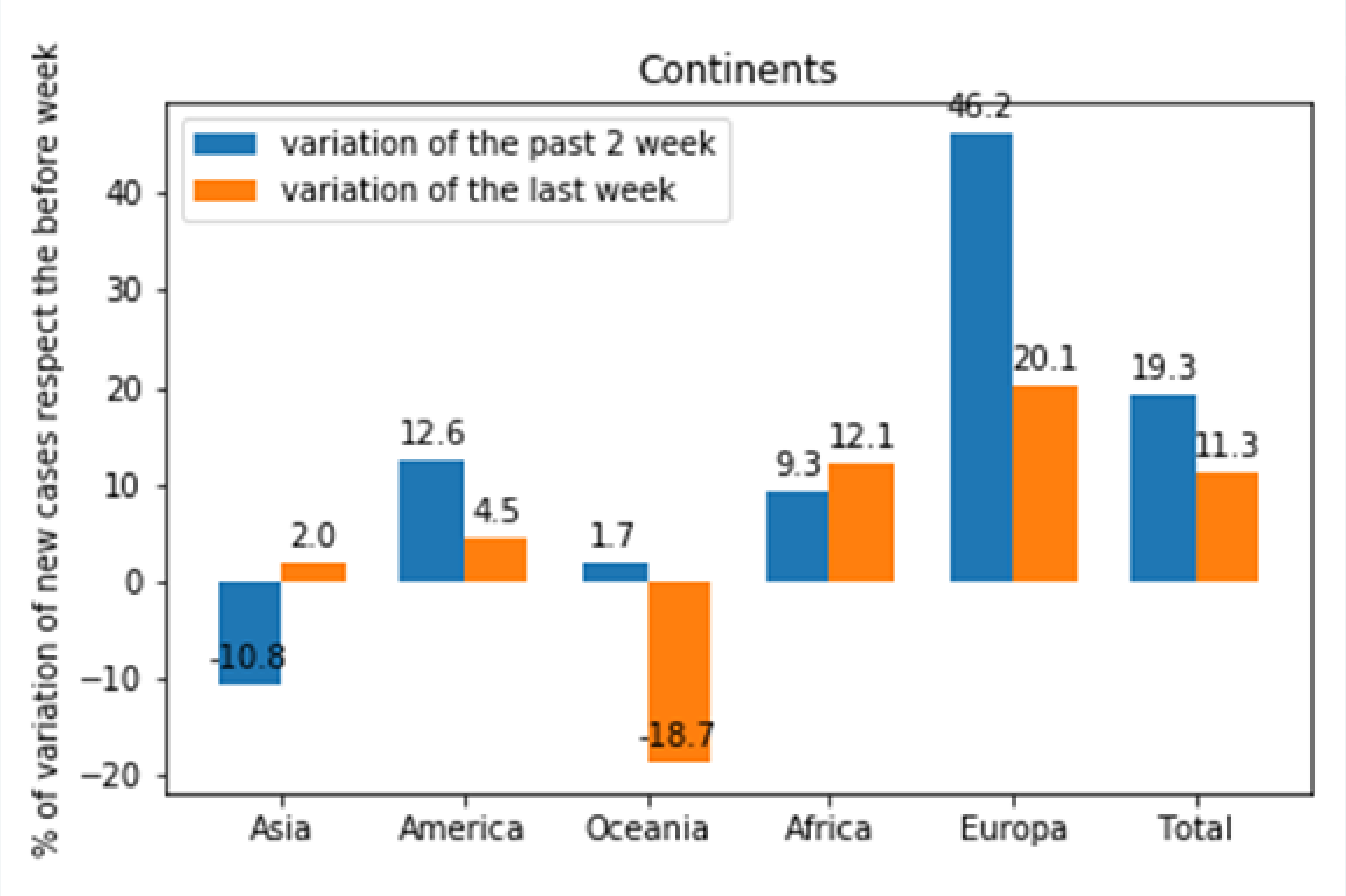


Linea del tiempo de casos por continentes y muertes totales.



Continent	New cases past 2 week	New cases past week	New cases last week	variation of the past week	variation of the week
Asia	582871	519926	530373	-10.8	2.0
America	847479	953977	997020	12.6	4.5
Oceania	35447	36048	29321	1.7	-18.7
Africa	187271	204700	229434	9.3	12.1
Europa	957913	1399990	1681354	46.2	20.1
Total	2610981	3114641	3467502	19.3	11.3

Casos Confirmados en las ultimas semanas.



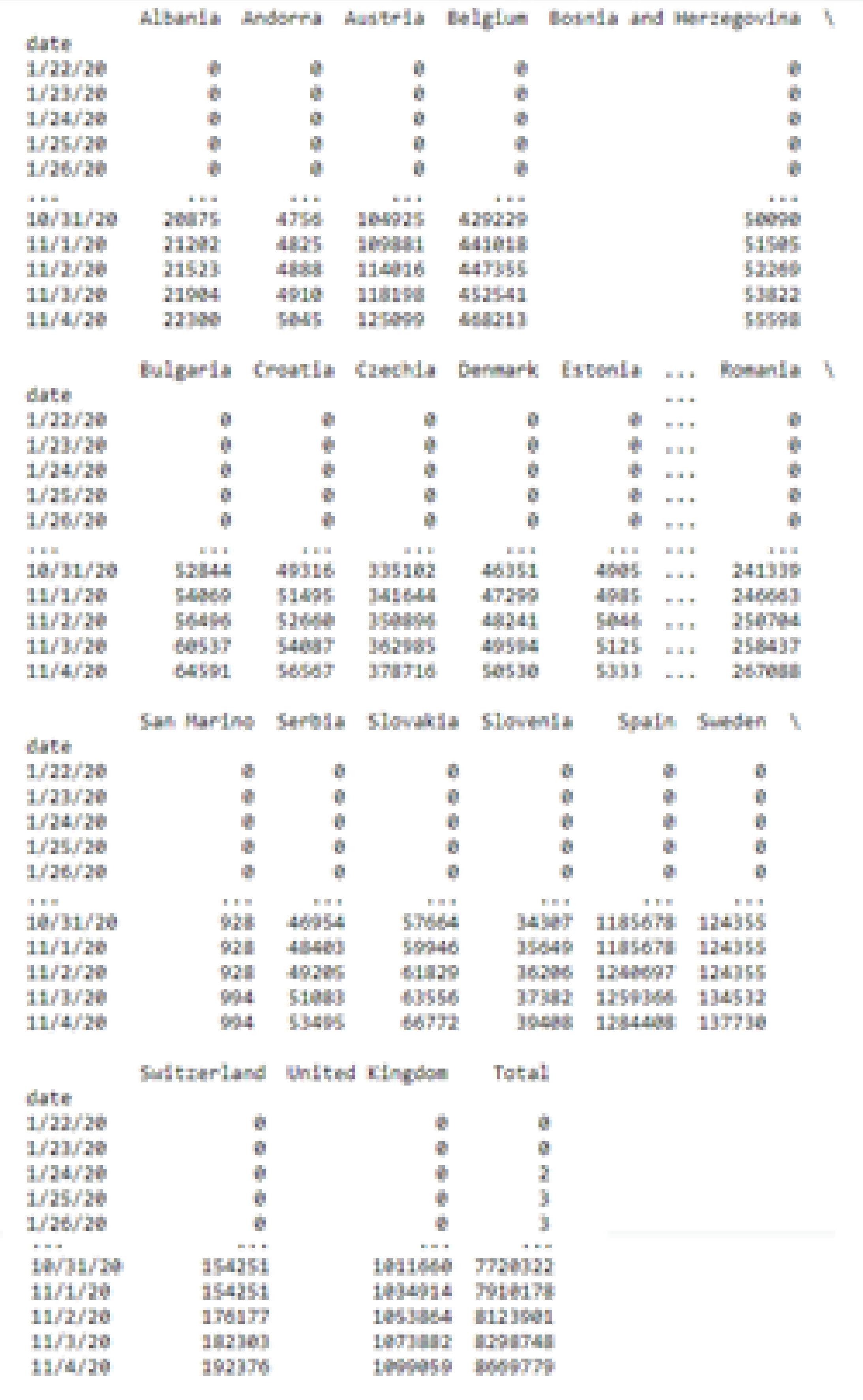
Variaciones semanales para los distintos continentes.

Choice of study region:

As seen in the course of the last week and in the penultimate Europe increased 46.2% and 20.1% of new cases compared to the previous weeks. Due to the increase in cases, it was decided to focus this analysis in Europe. Other

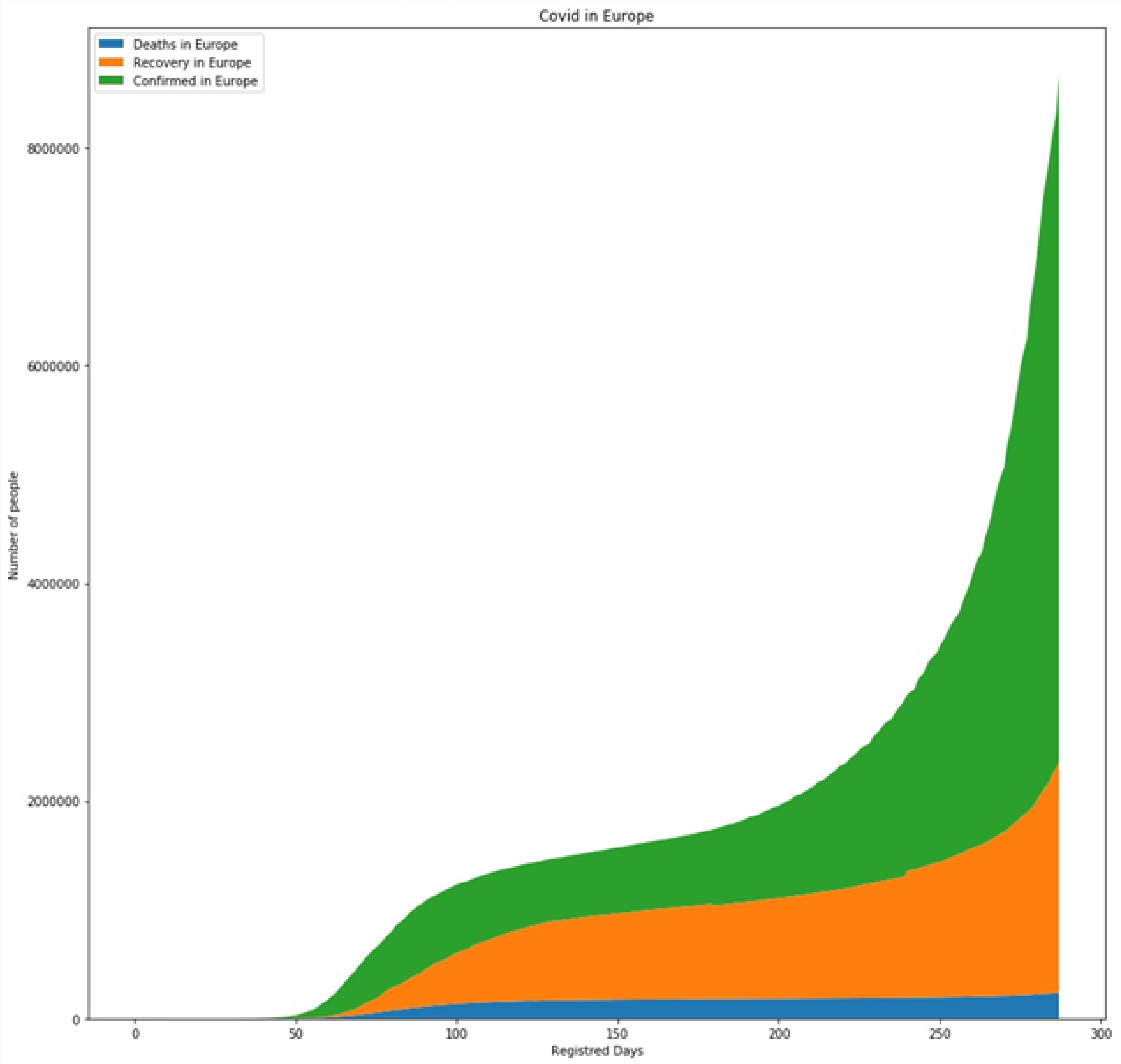
reasons why the choice are detailed below:

- The quality of the data we have access to is good.
- Largest population in the same region.
- Average age values for similar countries.
- Large number of records per unit of geographic area.



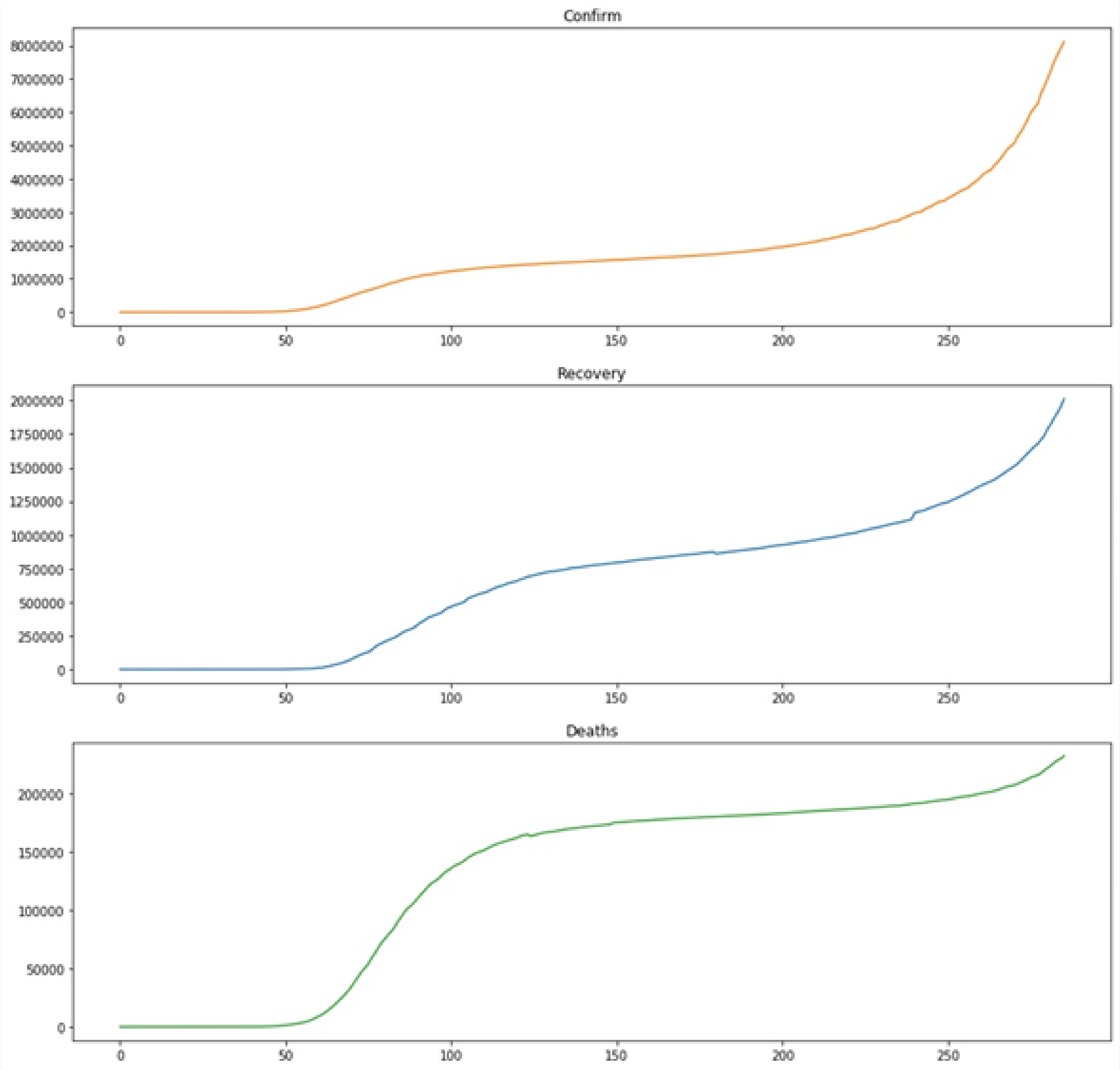
```
          Albania  Andorra  Austria  Belgium  Bosnia and Herzegovina  \
date
1/22/20         0        0        0        0                       0
1/23/20         0        0        0        0                       0
1/24/20         0        0        0        0                       0
1/25/20         0        0        0        0                       0
1/26/20         0        0        0        0                       0
...           ...      ...      ...      ...                     ...
10/31/20    20875     4756   104925   429229                   50090
11/1/20     21202     4825   109881   441018                   51505
11/2/20     21523     4888   114016   447355                   52269
11/3/20     21904     4910   118198   452541                   53822
11/4/20     22300     5045   125099   468213                   55598

          Bulgaria  Croatia  Czechia  Denmark  Estonia  ...  Romania  \
date                                                    ...
1/22/20          0        0        0        0        0  ...        0
1/23/20          0        0        0        0        0  ...        0
1/24/20          0        0        0        0        0  ...        0
1/25/20          0        0        0        0        0  ...        0
1/26/20          0        0        0        0        0  ...        0
...            ...      ...      ...      ...      ...  ...      ...
10/31/20     52844    49316   335102    46351     4905  ...   241339
11/1/20      54069    51495   341644    47299     4985  ...   246663
11/2/20      56496    52660   350896    48241     5046  ...   250704
11/3/20      60537    54087   362985    49594     5125  ...   258437
11/4/20      64591    56567   378716    50530     5333  ...   267088

          San Marino  Serbia  Slovakia  Slovenia    Spain  Sweden  \
date
1/22/20            0       0         0         0        0       0
1/23/20            0       0         0         0        0       0
1/24/20            0       0         0         0        0       0
1/25/20            0       0         0         0        0       0
1/26/20            0       0         0         0        0       0
...              ...     ...       ...       ...      ...     ...
10/31/20         928   46954     57664     34307  1185678  124355
11/1/20          928   48403     59946     35649  1185678  124355
11/2/20          928   49205     61829     36206  1240697  124355
11/3/20          994   51083     63556     37382  1259366  134532
11/4/20          994   53495     66772     39408  1284408  137730

          Switzerland  United Kingdom    Total
date
1/22/20             0               0        0
1/23/20             0               0        0
1/24/20             0               0        2
1/25/20             0               0        3
1/26/20             0               0        3
...               ...             ...      ...
10/31/20       154251         1011660  7720322
11/1/20        154251         1034914  7910178
11/2/20        176177         1053864  8123901
11/3/20        182303         1073882  8298748
11/4/20        192376         1099059  8669779
```

Table of confirmed in Europe:



Europe Timeline (Confirmed, Recovered, Dead

Europe Timeline Analysis:
As can be seen in the upper timeline, the evolution of infections in the last few weeks has increased in an exponential way. Now we will try to find patterns between the confirmed, dead and recovered timelines that allow me to correlate them by separating their graphs.



Separando las lineas de tiempo:

It can be seen that the dead and recovered curve resembles the confirmed curve after a few days. We will try to find how many days are necessary for both curves to match.

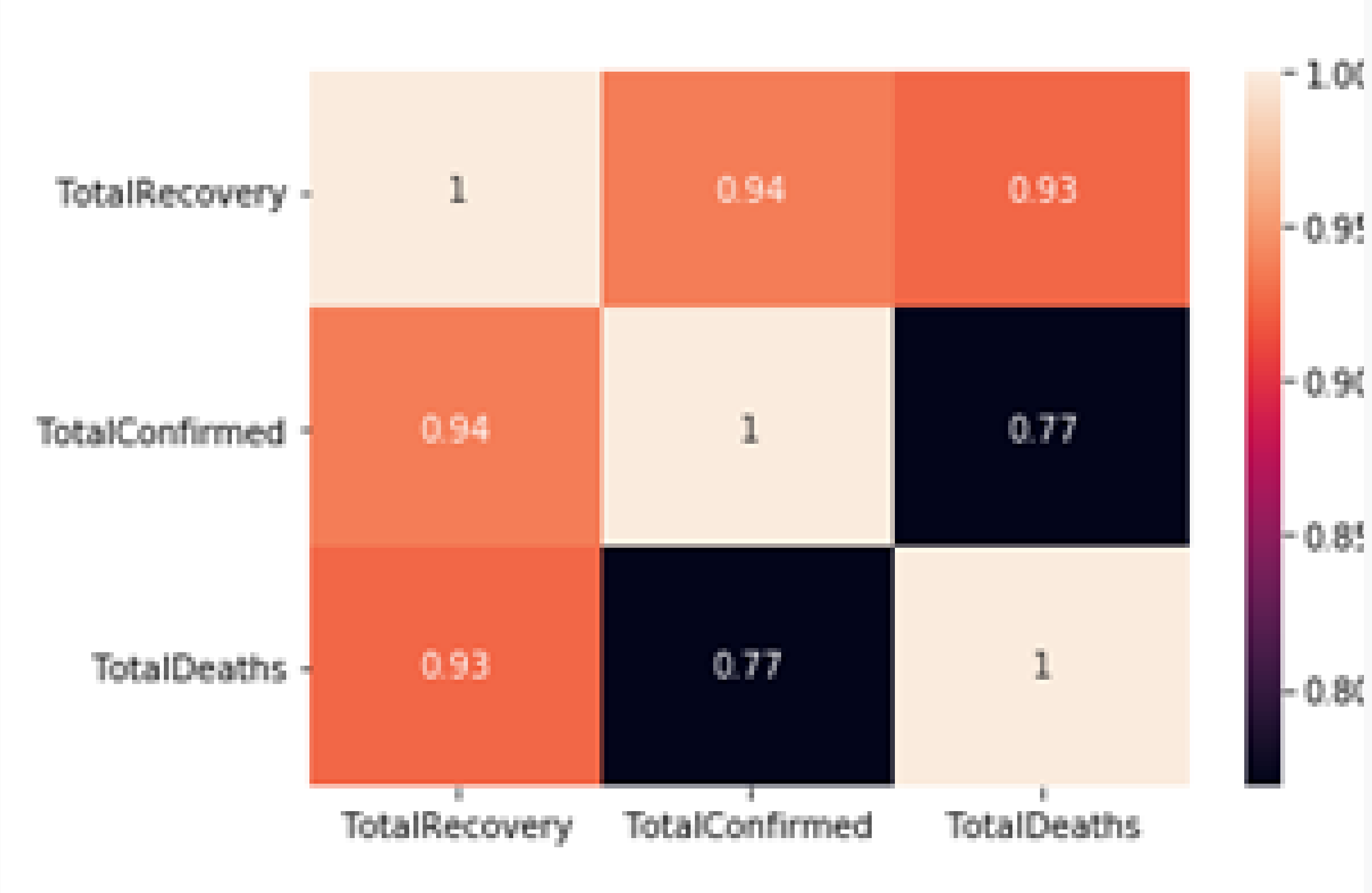


Gráfico de Pearson: curvas reales

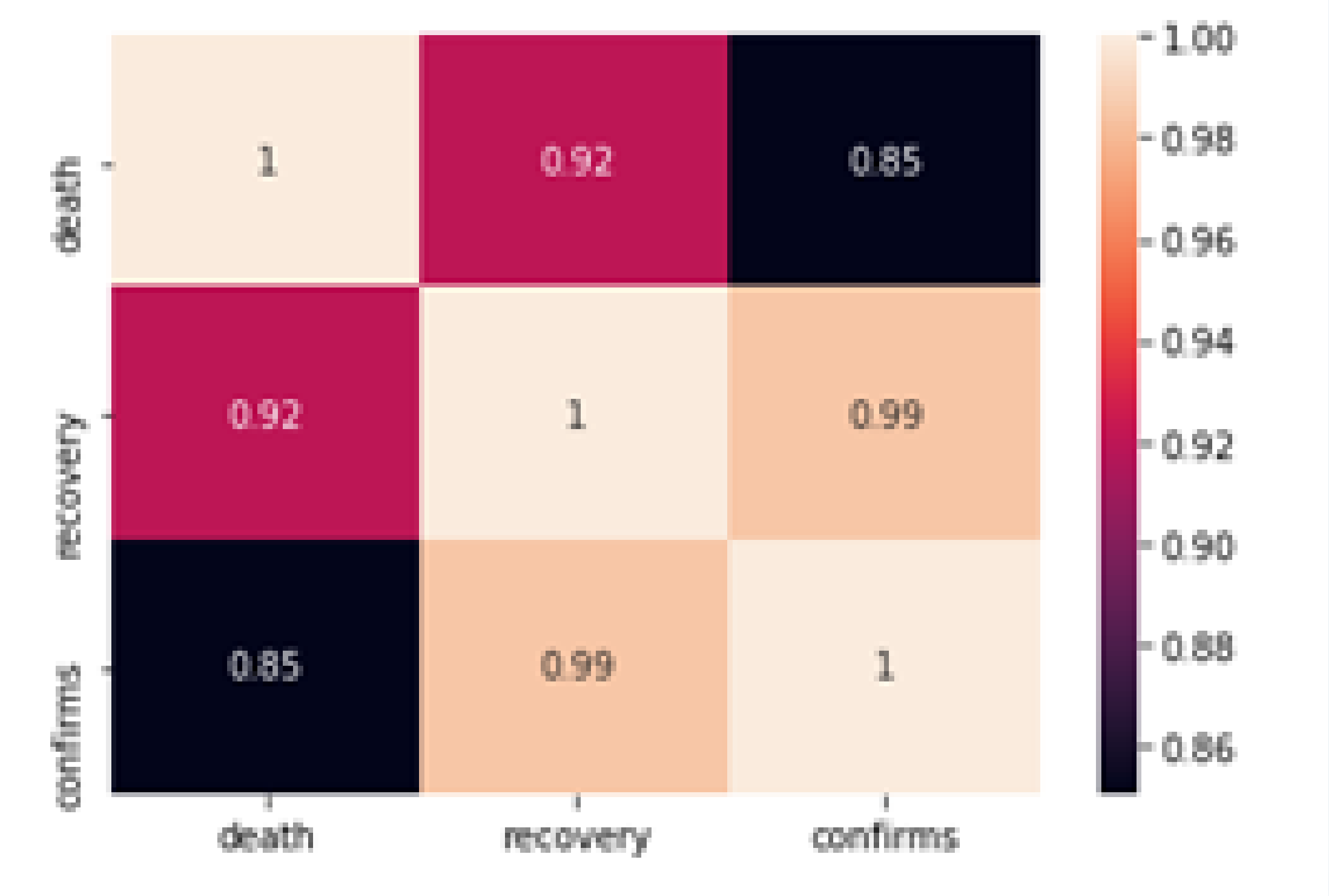
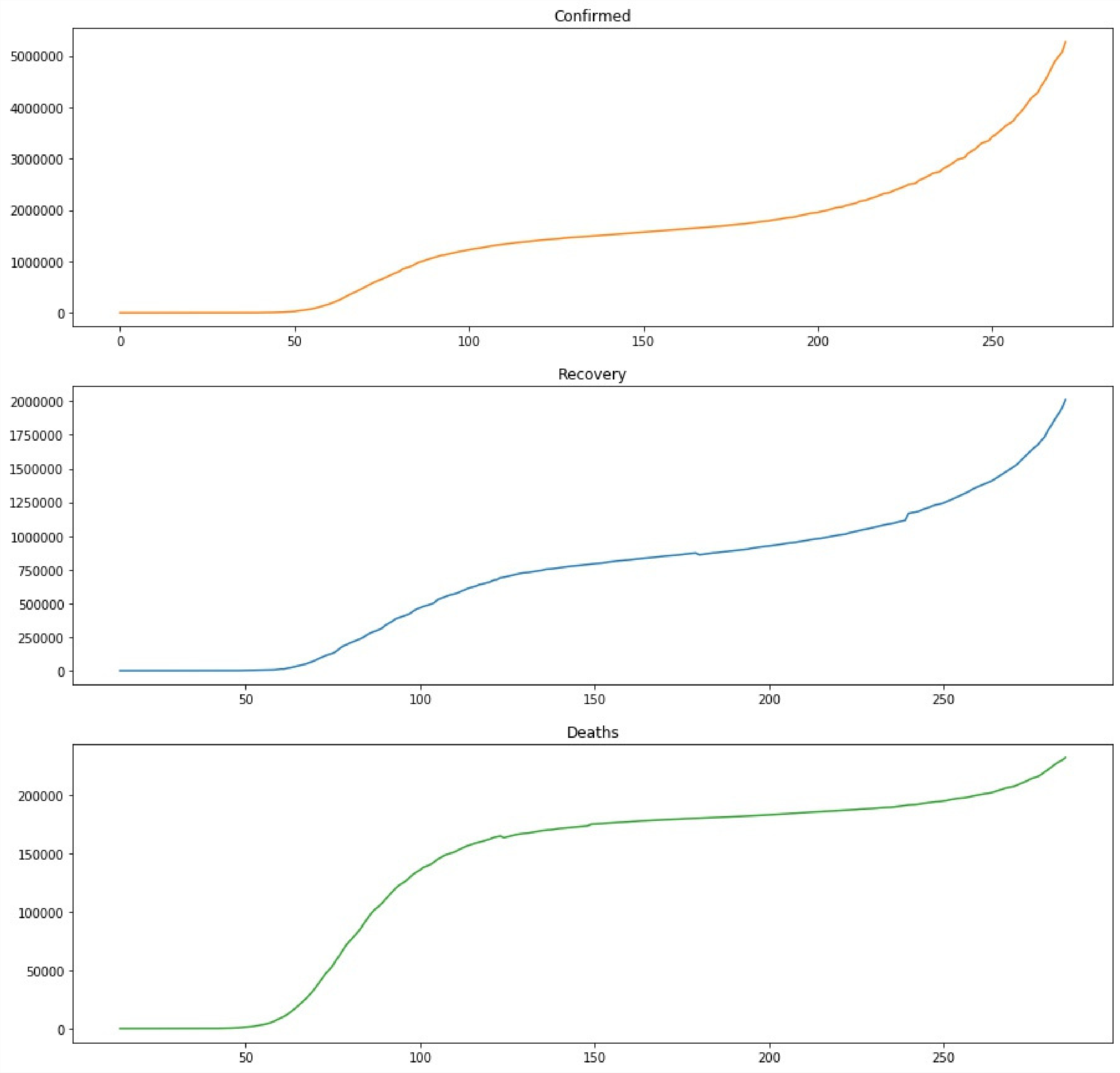


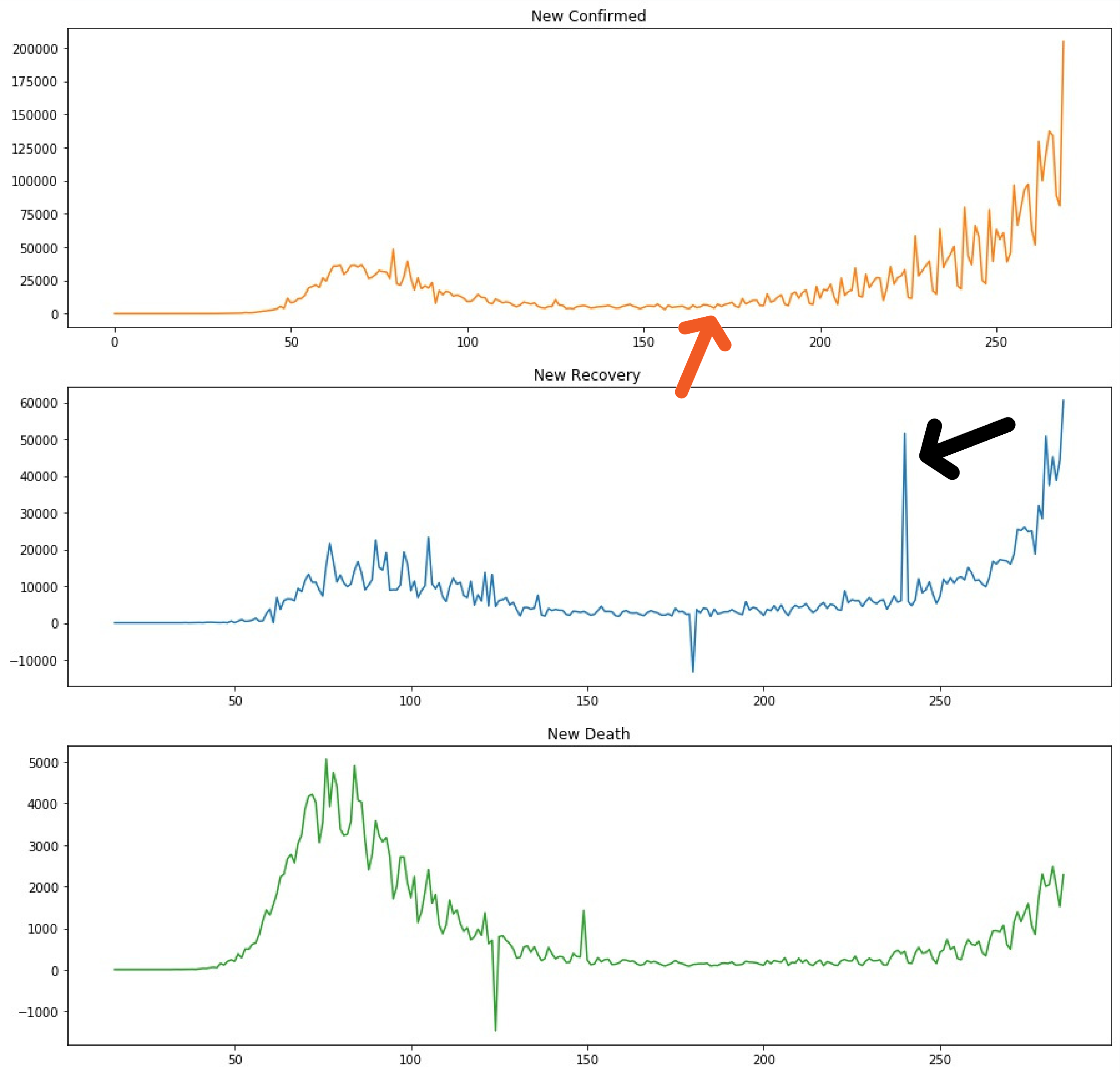
Gráfico de Pearson: curvas de confirmados atrazados 14 dias

Hypothesis analysis of offset days in the curves:
After an analysis and applying Pearson's linear correlation statistical techniques, it was determined that after 14 days the correlation improves between the confirmed and dead curve of
0.77 to 0.85. Regarding confirmed and recovered from 0.93 to 0.99. Both give a very good correlation and allow us to advance with our hypotheses.

Updating our curves
Adapting the curves to our hypothesis, we generate a good correlation to analyze what is happening in recent weeks and estimate from when the second wave of infections began in Europe and what its current growth is. In this way we could also predict the dead and confirmed for
the next few weeks



otal records



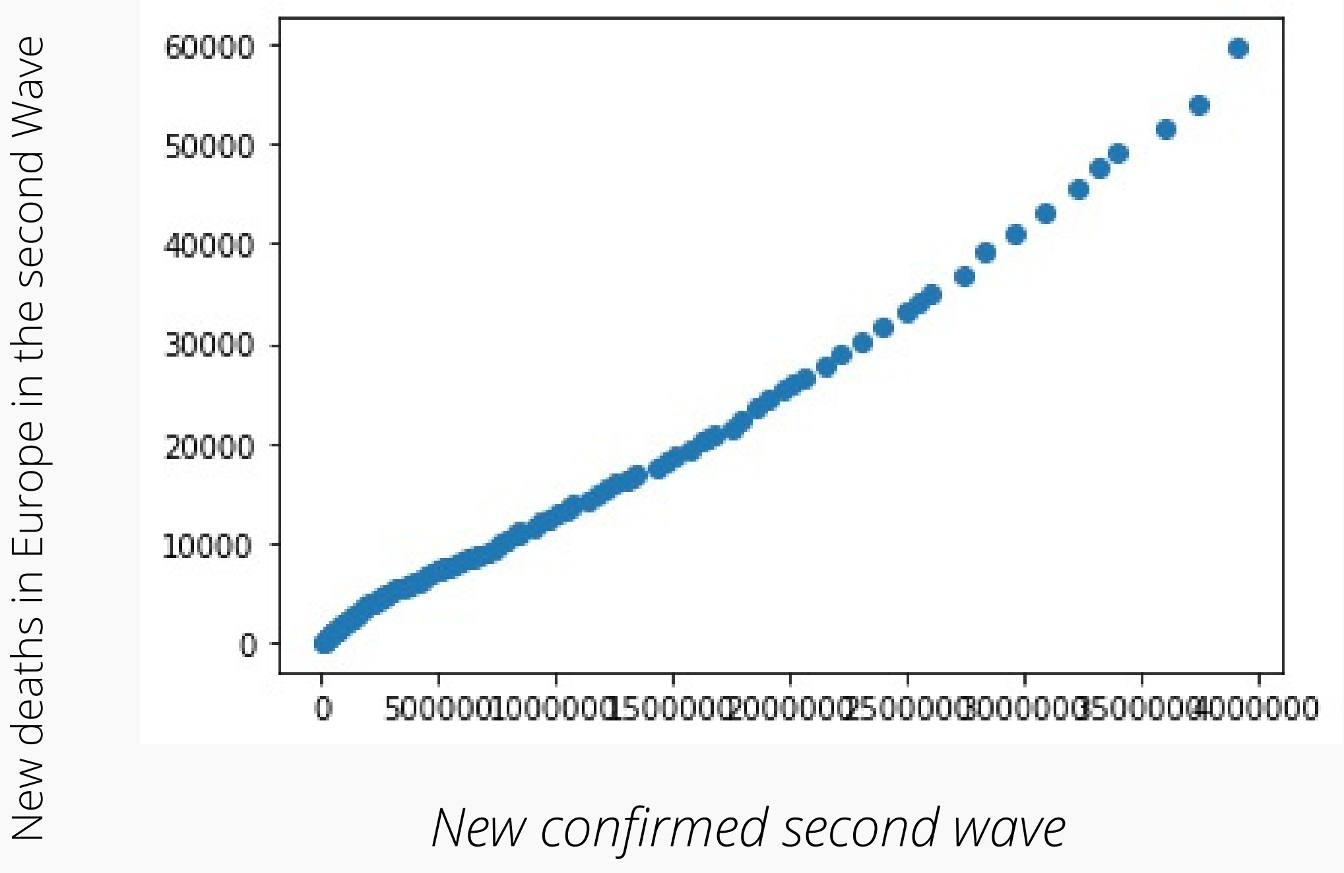
New records per day

Visualization of the upper graphics

As you can see, Europe is going through a second wave that has started from day 170 registered with a red arrow. As can also be seen, in the previous graph of new records a peak is observed for those recovered, this indicates that a country updated its recovered that day (black arrow). Romania was the country that in the report of 09/18/2020 all its recovered from covid 19: 45654 recovered

Timeless second wave analysis

We analyze the correlation that exists between confirmed and timeless deaths of the second wave of covid 19 in Europe and we try to find a good model that allows us to predict what will happen in the next two weeks



New confirmed second wave

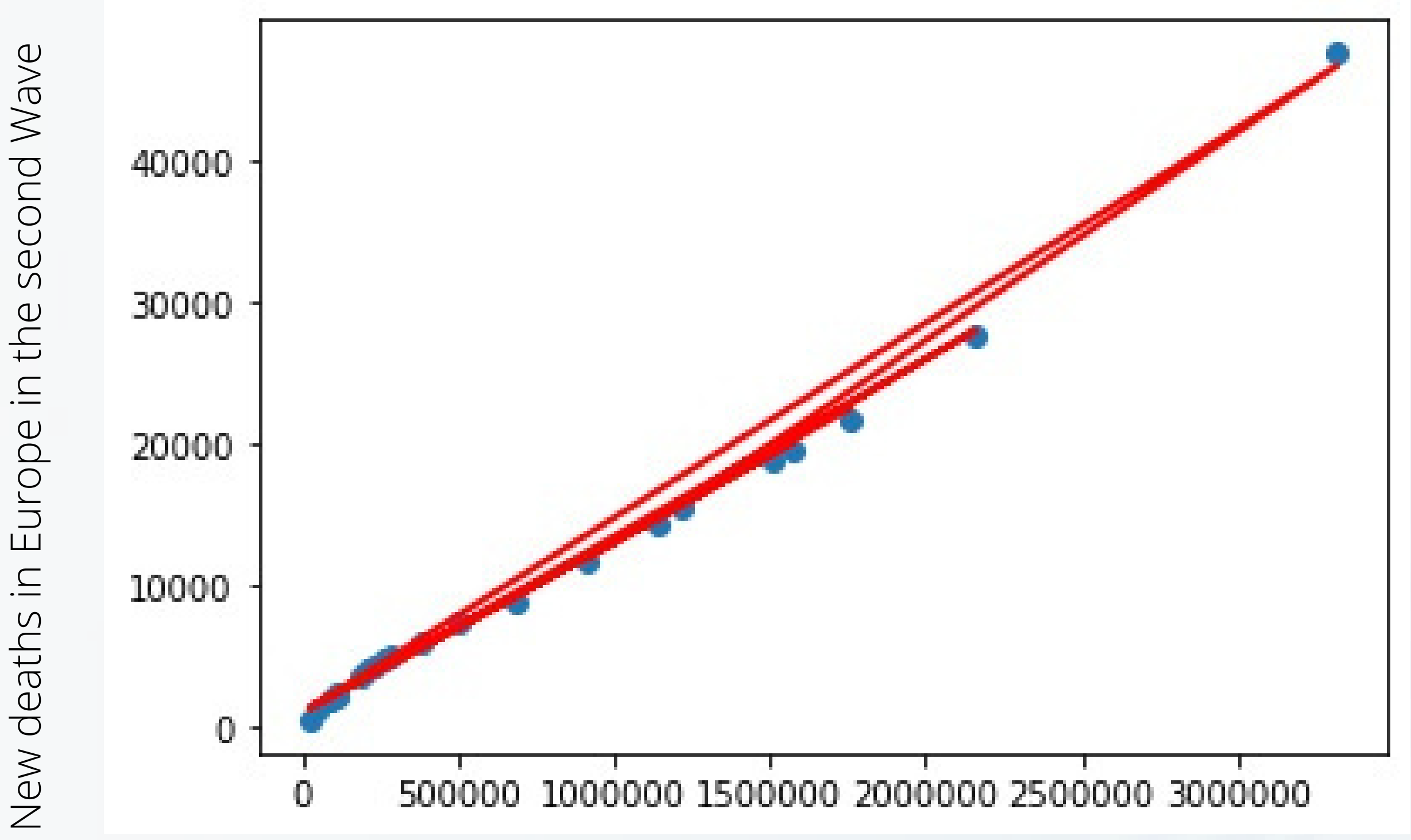
Trying to find a good model that fits. It is seen that the curve fits a linear regression but a good prediction could also be found with a higher degree polynomial regression model. And it will be adjusted with the MSE metric

$$MSE = \frac{1}{n} \Sigma \underbrace{\left(y - \widehat{y} \right)^2}_{\text{The square of the difference between actual and predicted}}$$

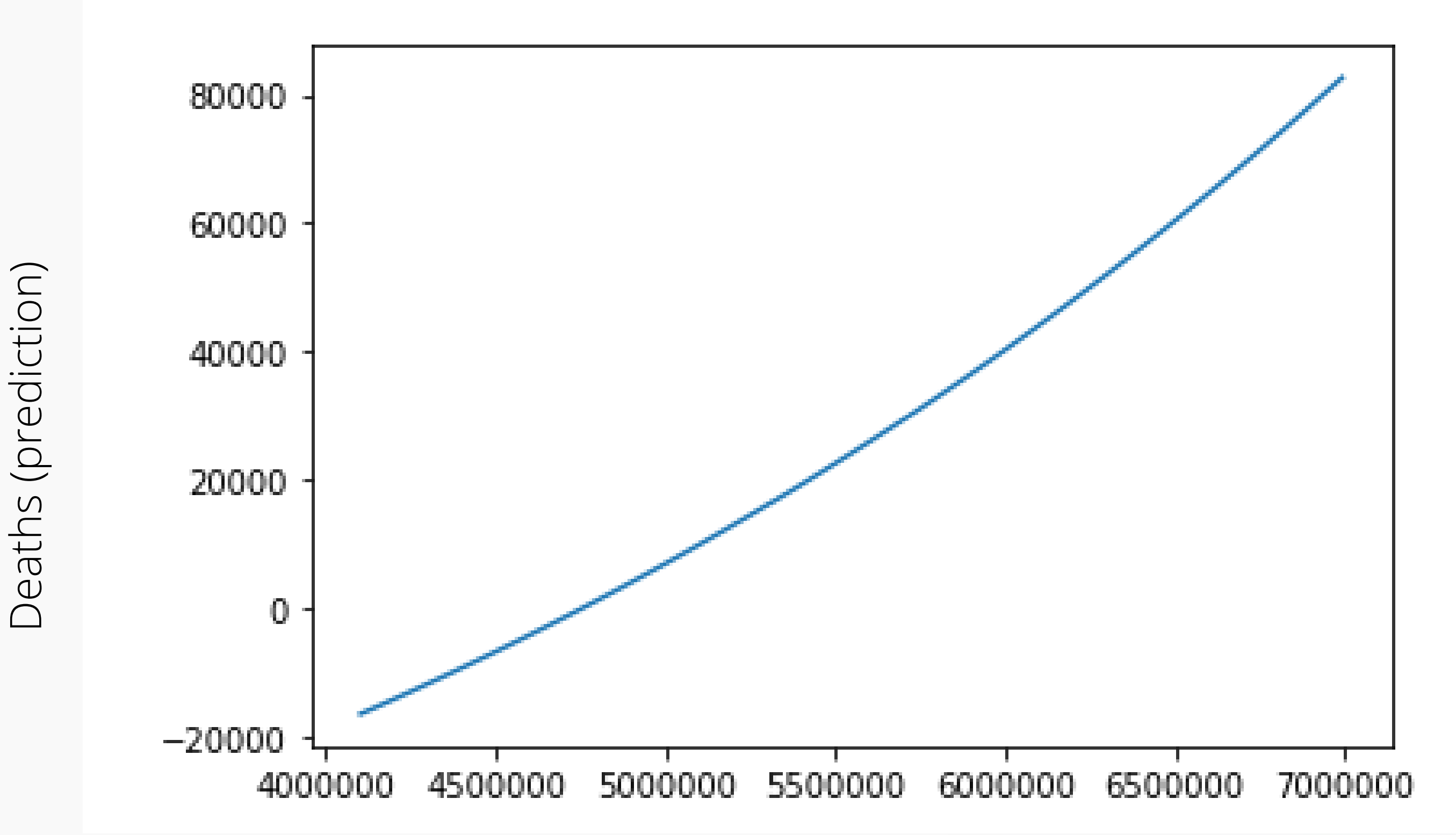
y = current
y ^ =^prediction
n = number of samples

The model was trained with polynomial regression for degrees 2,3,4 (higher polynomial degrees the model will be in overfitting). To evaluate the results, we used the metric of the lowest mean square error (mse) and the highest prediction score, the accuracy was determined and the model that best adjusted to reality was decided based on these two metrics.

Fitted model values
Degree of polynomial 3 Value
Value of coeficience a
[[0.00000000e+00 1.20179868e-02 -3.82819980e-10 2.81025203e-16]]
Value of intersection b
[989.91
precition of the model
R= 0.9991141076431113
Mean Square Error
244724.0



New confirmed second wave (test values vs prediction values)

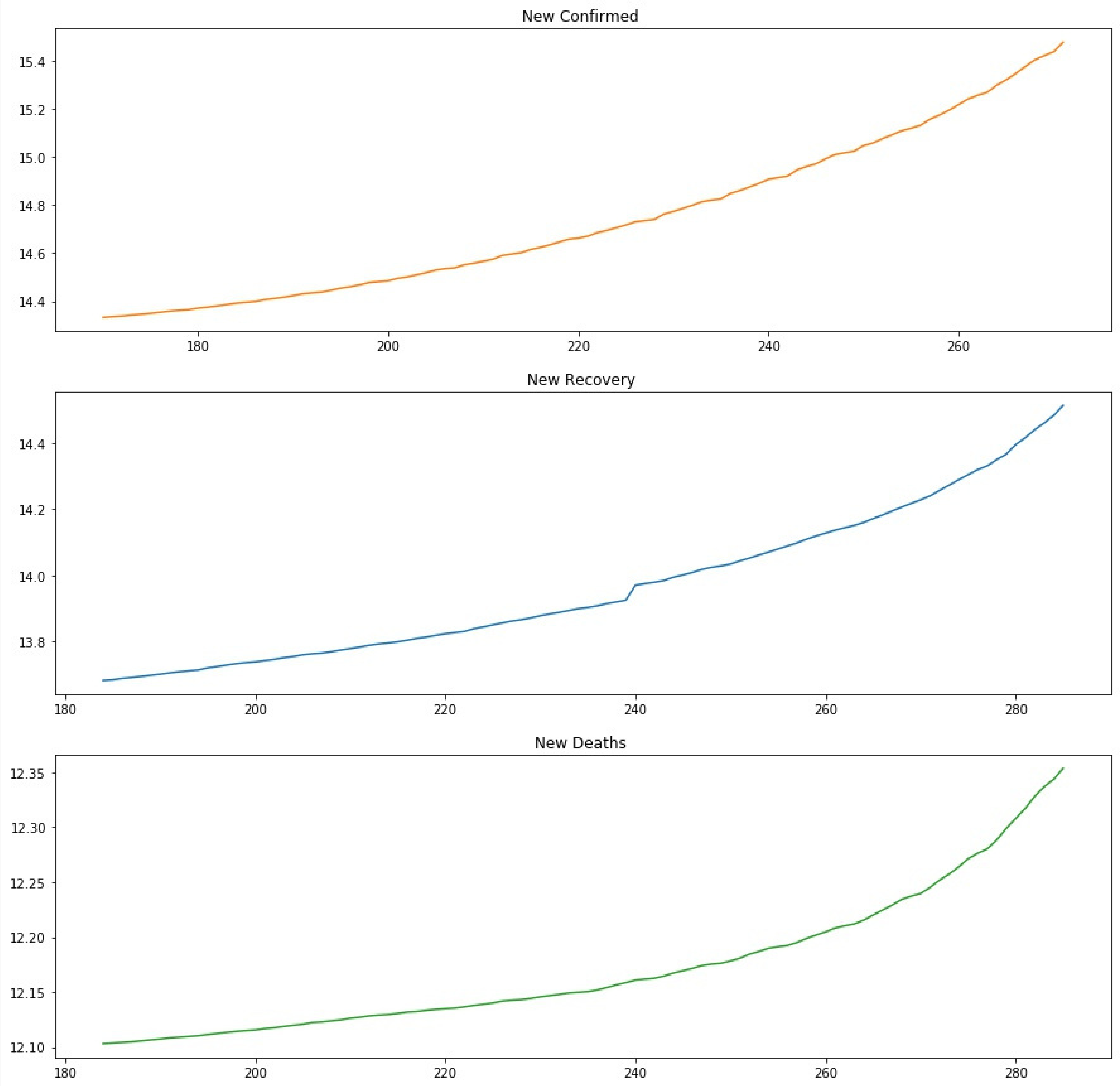


Cases confirmed of the second wave actually and deaths for the next weeks

Conclusion of this prediction analysis:
counting the current confirmed of the second wave with respect to the current confirmed 80,000 new deaths are expected in Europe in the coming weeks corresponding to those infected recently.

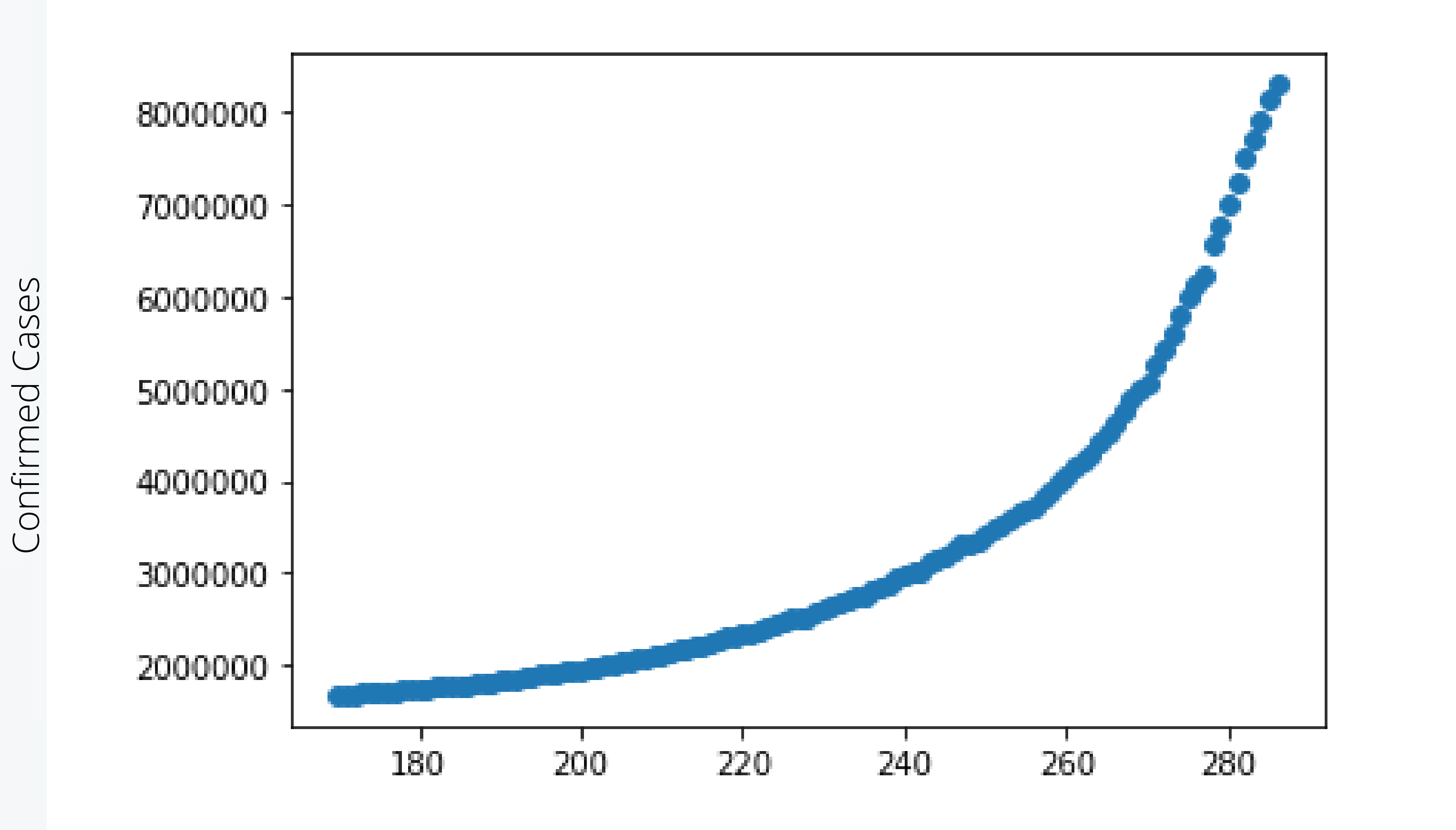
Time estimate for the second wave:
In our next analysis we will not focus on a temporal comparison comparing the deaths and confirmed deaths of the second wave in Europe and thus be able to estimate in the coming weeks how these indicators will advance.

I apply logarithms to the graphs of the timeline of the second wave, offsetting the confirmed ones by 14 days so that it correlates with dead and recovered and thus know what type of method to use.

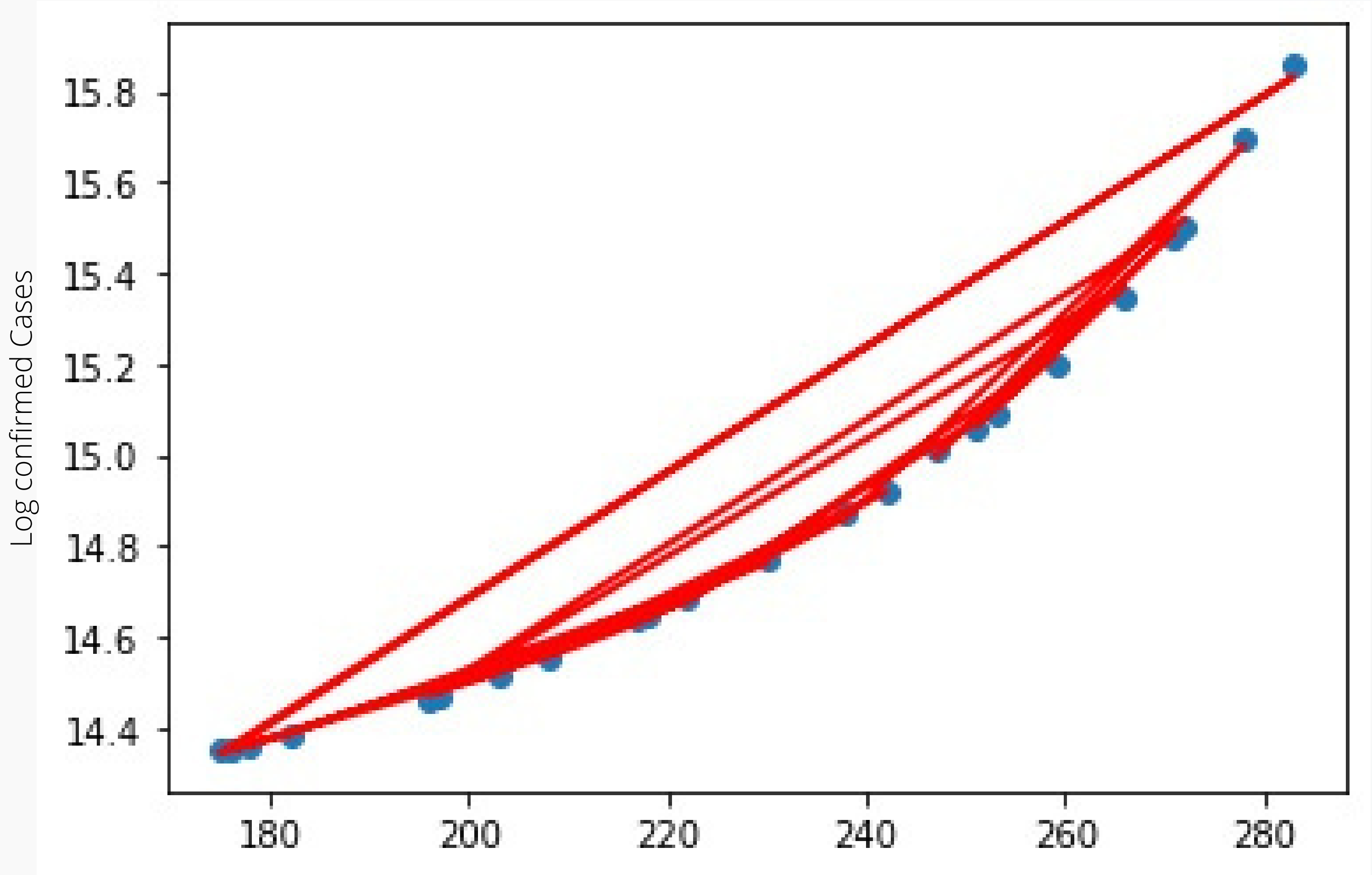


As can be seen in the upper graph, Europe faces a second wave with exponential growth, the first 30 days of this period exponential growth was constant. This means that in the last days of the second wave, growth accelerated compared to what was seen in the first part
of the second wave, with this growth, if it continues like this during the next 14 days, the number of new infections can be estimated. and there will be deaths in 14 days, we are going to evaluate this premise.

Confirmed cases in Europe that occurred since the second wave began (which we intuit that it begins on the registered day n ° 170)



Days Registed of the second wave

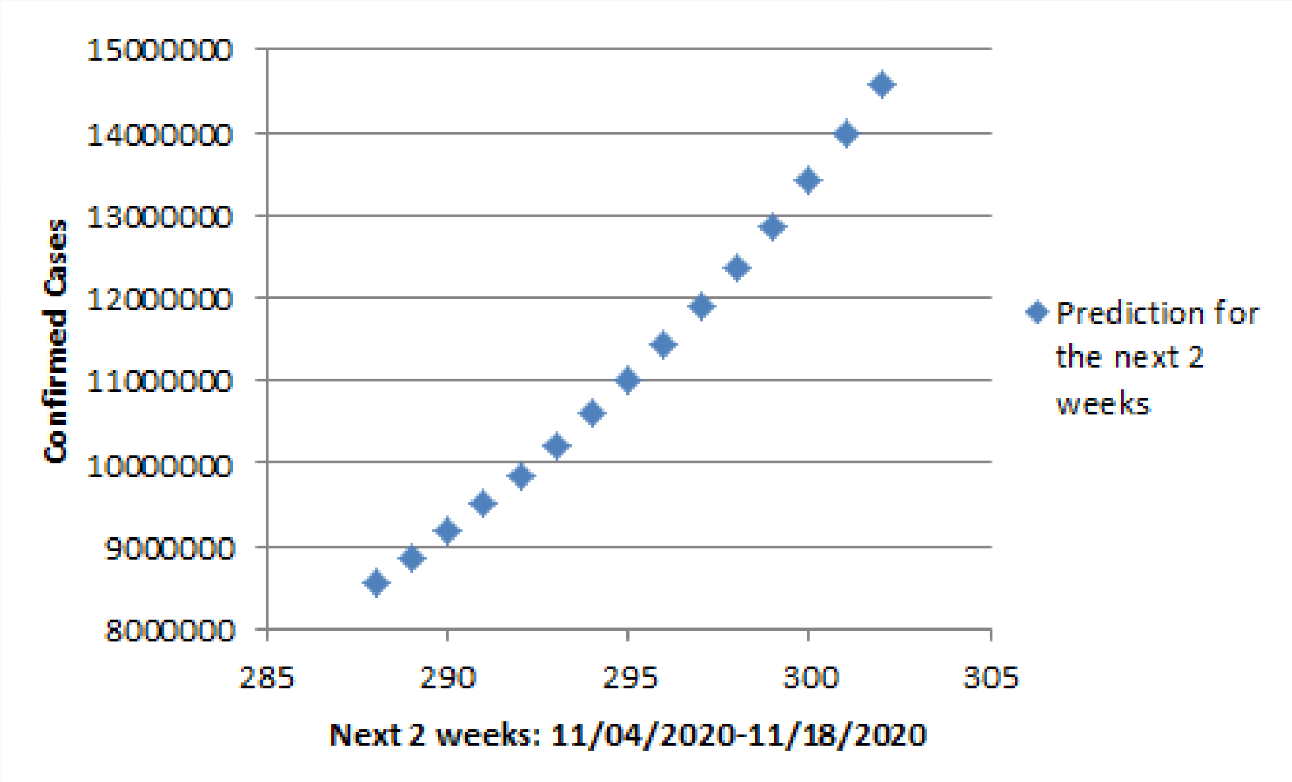


Days Registed of the second wave (values of test vs prediction values)

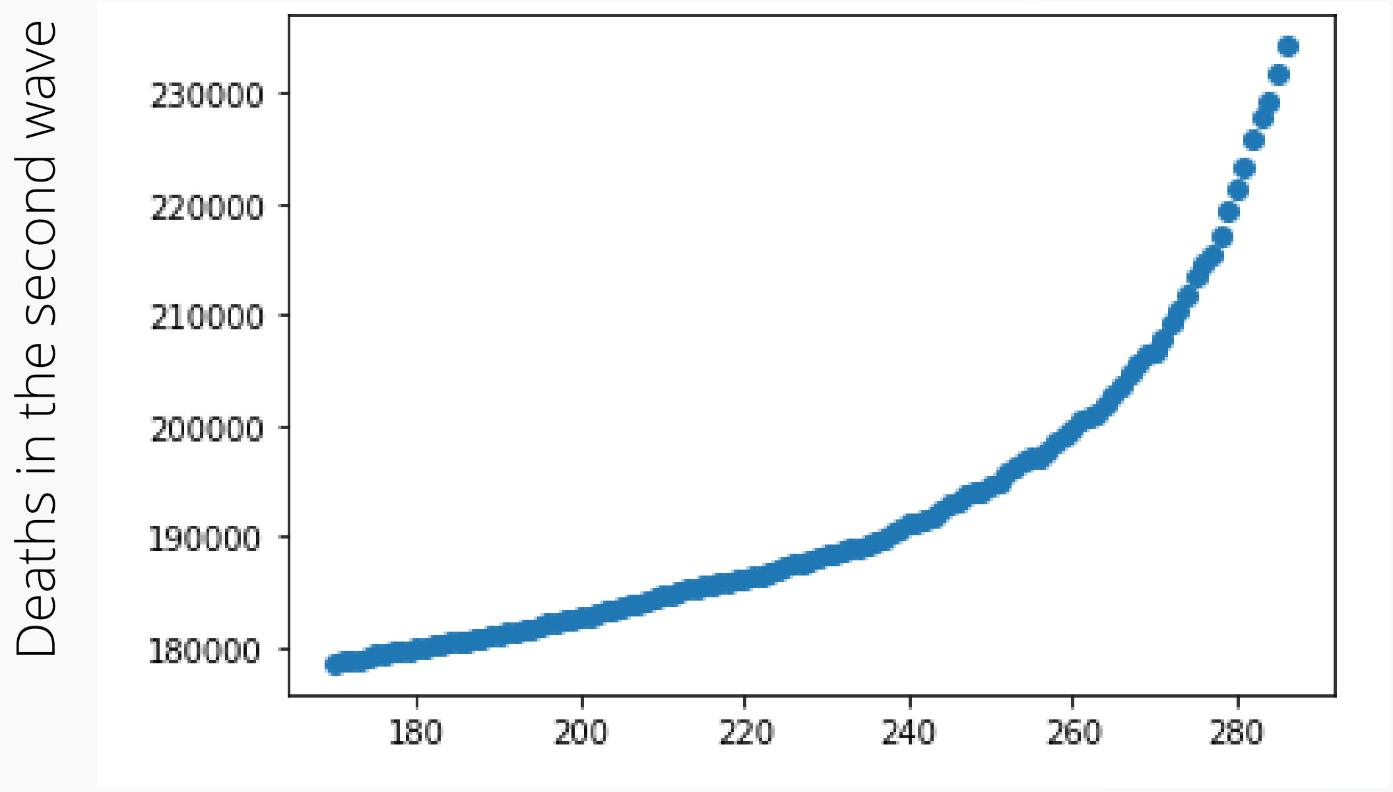
Data for the model of polinomial regresion
Polinomial degree 3
value of coeficient a
[0.00000000e+00 1.00910618e-01 -5.06494321e-04 9.04573811e-07]
value of interception
7.339632261454197
Model Precision
R= 0.9991997010148429
Mean Square Error
0.00025213659868412514

It can be seen that the predictor value for 11/4/20 is: 8585507.0 vs the real value: 8669779
error for this estimate: the error in this prediction was 0.98%, with this model we can predict in a short time with a low level of errorr

EEstimation of this analysis:
In two weeks it can be seen that confirmed cases in Europe will amount to 1,457,000, of which 6,000,000 will occur in the next two weeks

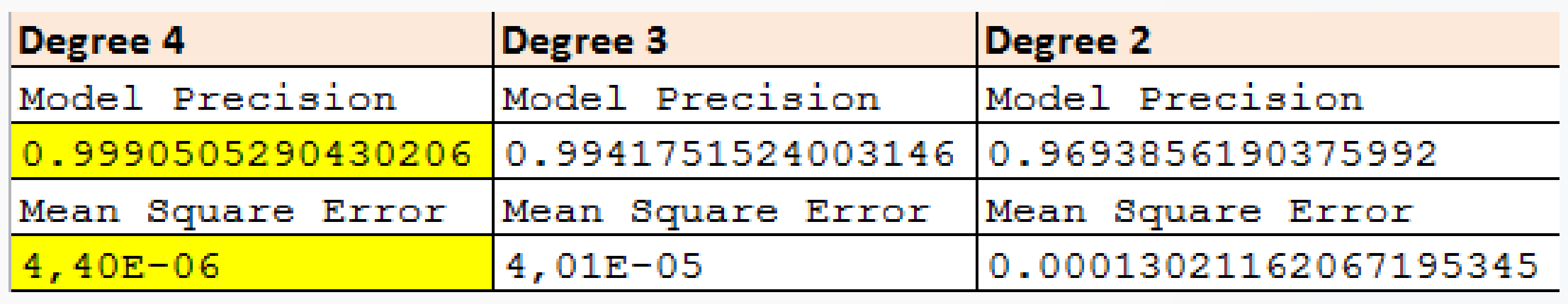


We will perform this temporal regression analysis again, comparing the deaths produced in the second wave as a function of days. and then estimate the number of deaths that will occur in the next two weeks



registered days of the second wave

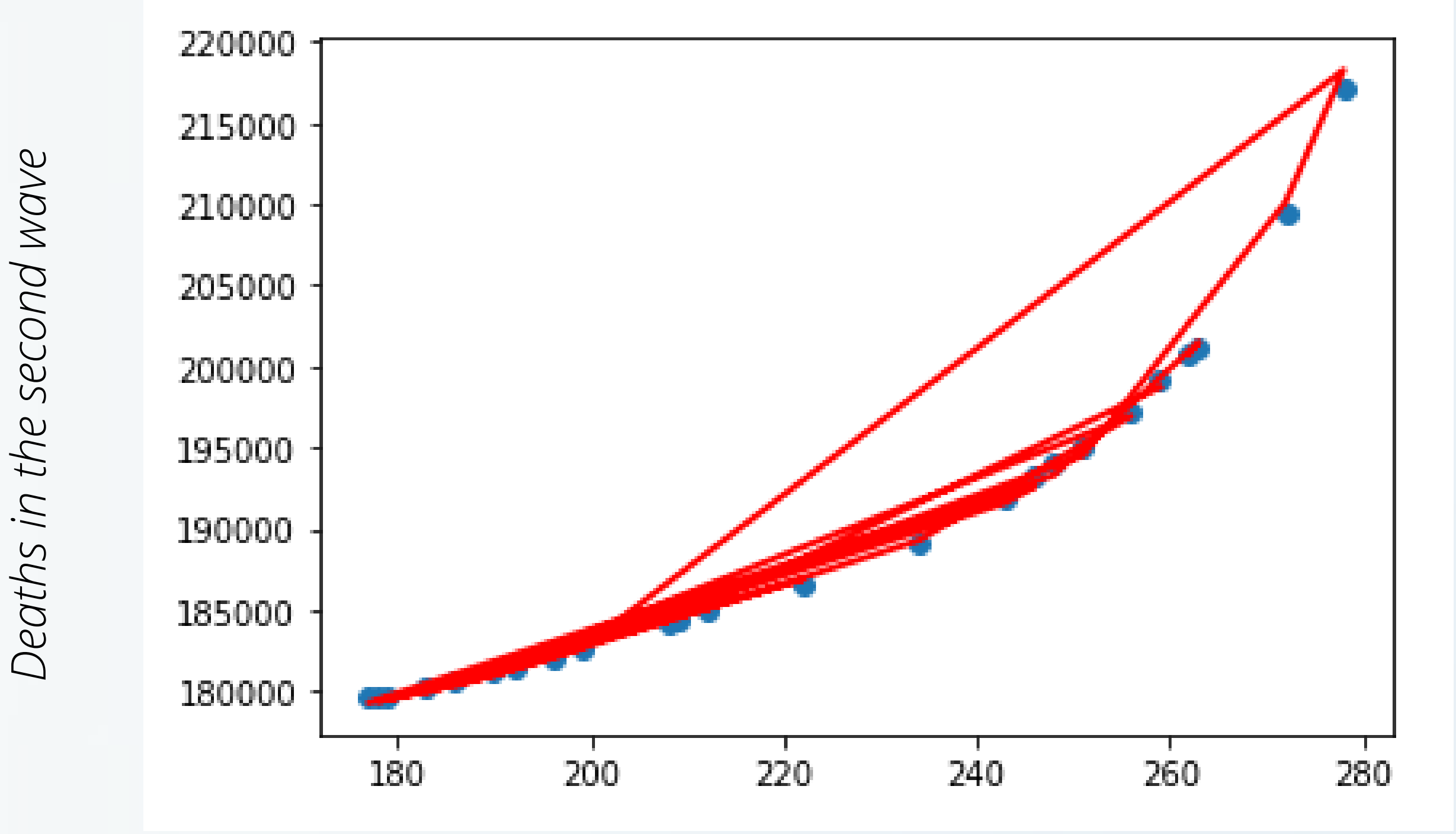
Currently the graph of deaths in the second wave shows the updating of Europe



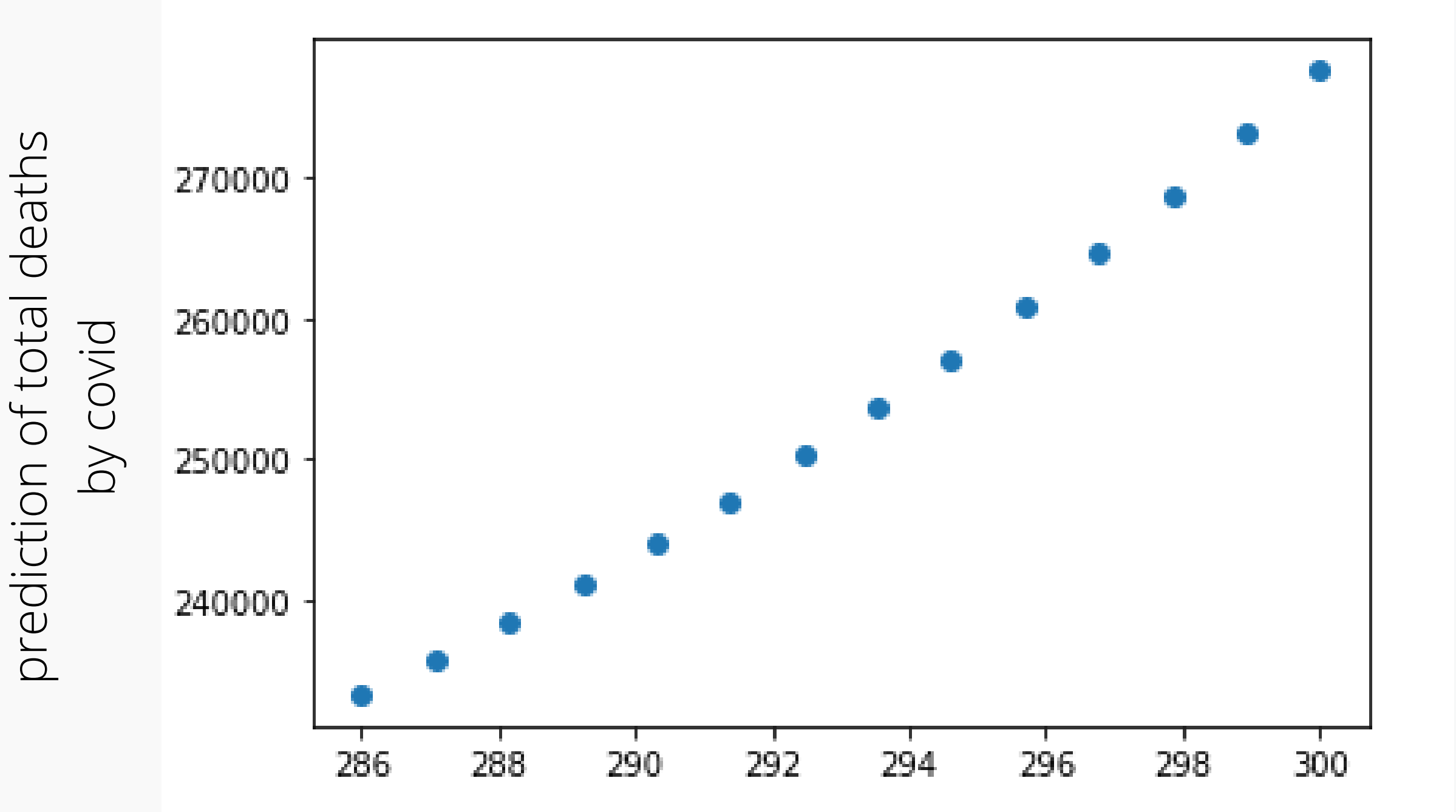
Degree 4	Degree 3	Degree 2
Model Precision	Model Precision	Model Precision
0.9990505290430206	0.9941751524003146	0.9693856190375992
Mean Square Error	Mean Square Error	Mean Square Error
4,40E-06	4,01E-05	0.00013021162067195345

valores de las metricas utilizadas en este caso

Data for the model of polinomial regresion
Polinomial Degree 4
value of coeficient a
[0.00000000e+00 -1.89184580e-01 1.36233896e-03
-4.32941660e-06
5.15048310e-09]
value of interception
21.85408841106984
Model Precision
R = 0.9990505290430206
Mean Square Error
4.400515965752392e-06



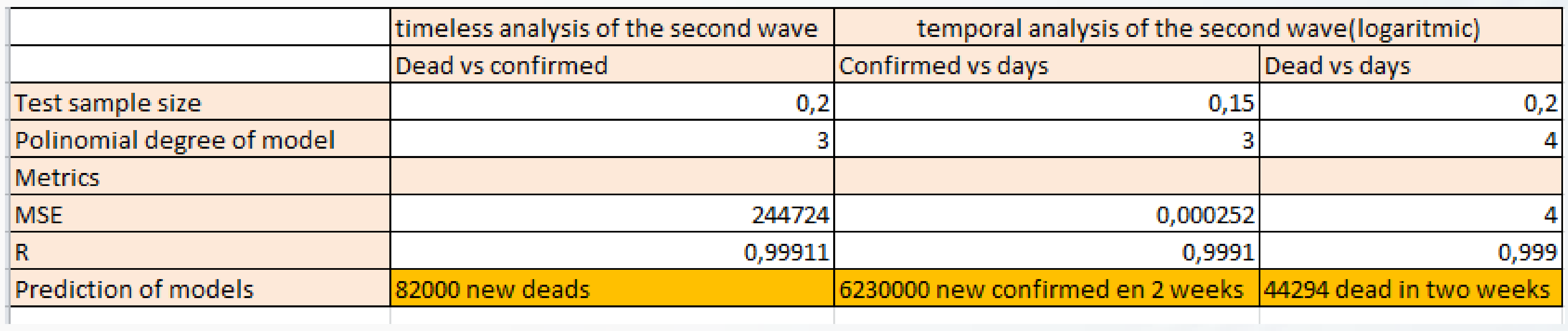
days recorded second wave



Days of the next two weeks

With respect to this prediction, 44294 deaths are expected in 2 weeks in Europe

Data conclusions:

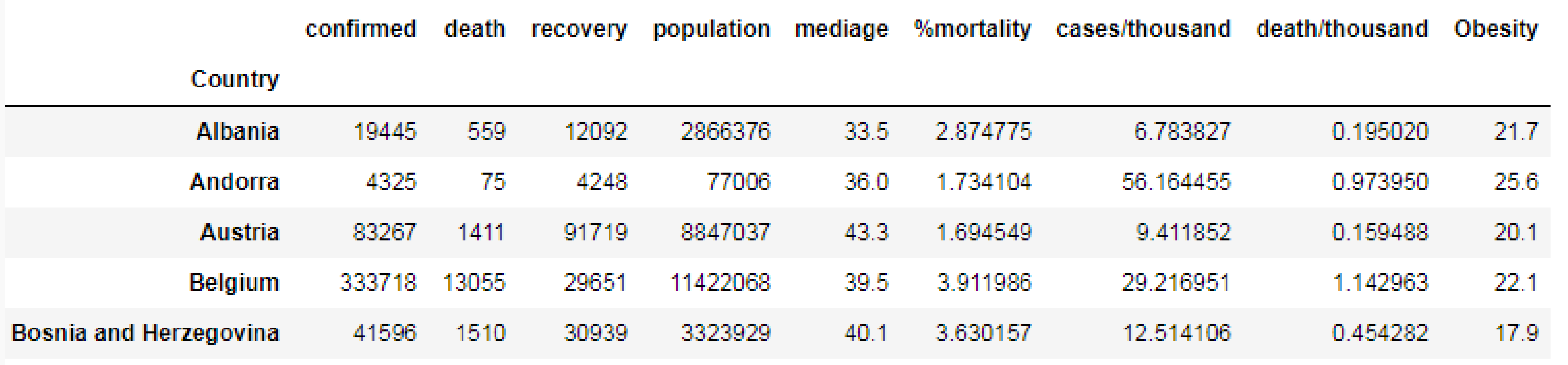


	timeless analysis of the second wave	temporal analysis of the second wave(logaritmic)	
	Dead vs confirmed	Confirmed vs days	Dead vs days
Test sample size	0,2	0,15	0,2
Polinomial degree of model	3	3	4
Metrics			
MSE	244724	0,000252	4
R	0,99911	0,9991	0,999
Prediction of models	82000 new deads	6230000 new confirmed en 2 weeks	44294 dead in two weeks

Conclusion
With regard to the two prediction models, it is expected that between 44,294 and 6,234,000 people will die from COVID in 2 weeks in Europe with a maximum of 78,800 deaths in the coming weeks, in view of the fact that the
contagions if a peak of contagions is not reached, the subsequent weeks will be more contagious and deadly. Despite having more tools the European Union to face the pandemic in this second wave seems to be a more intense
outbreak than in the first wave.

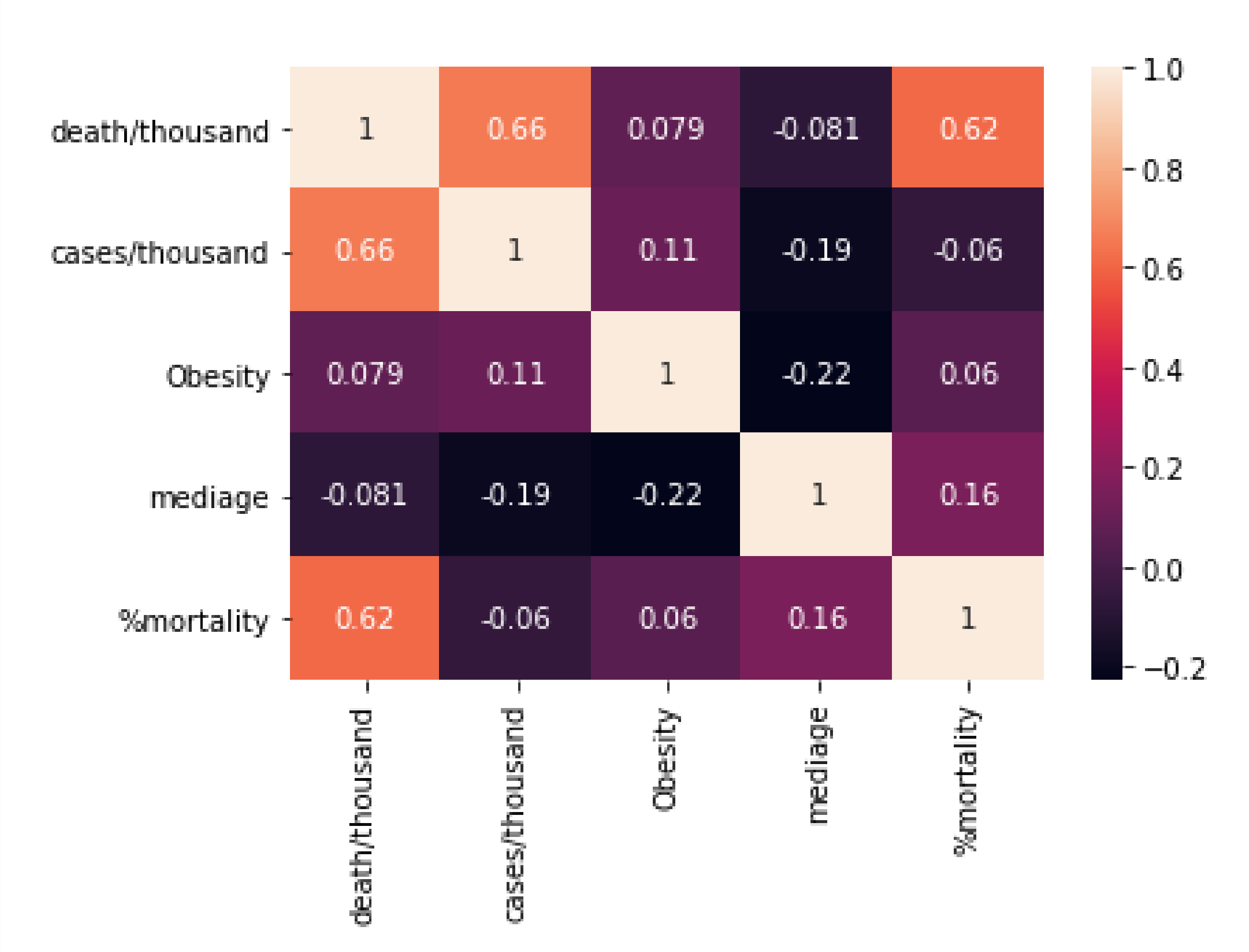
Now we will proceed to analyze mortality based on the characteristics of the population, mean age and obesity index, data obtained from the WHO and the World data bank, the treatment of the data is in the attached ipynb file

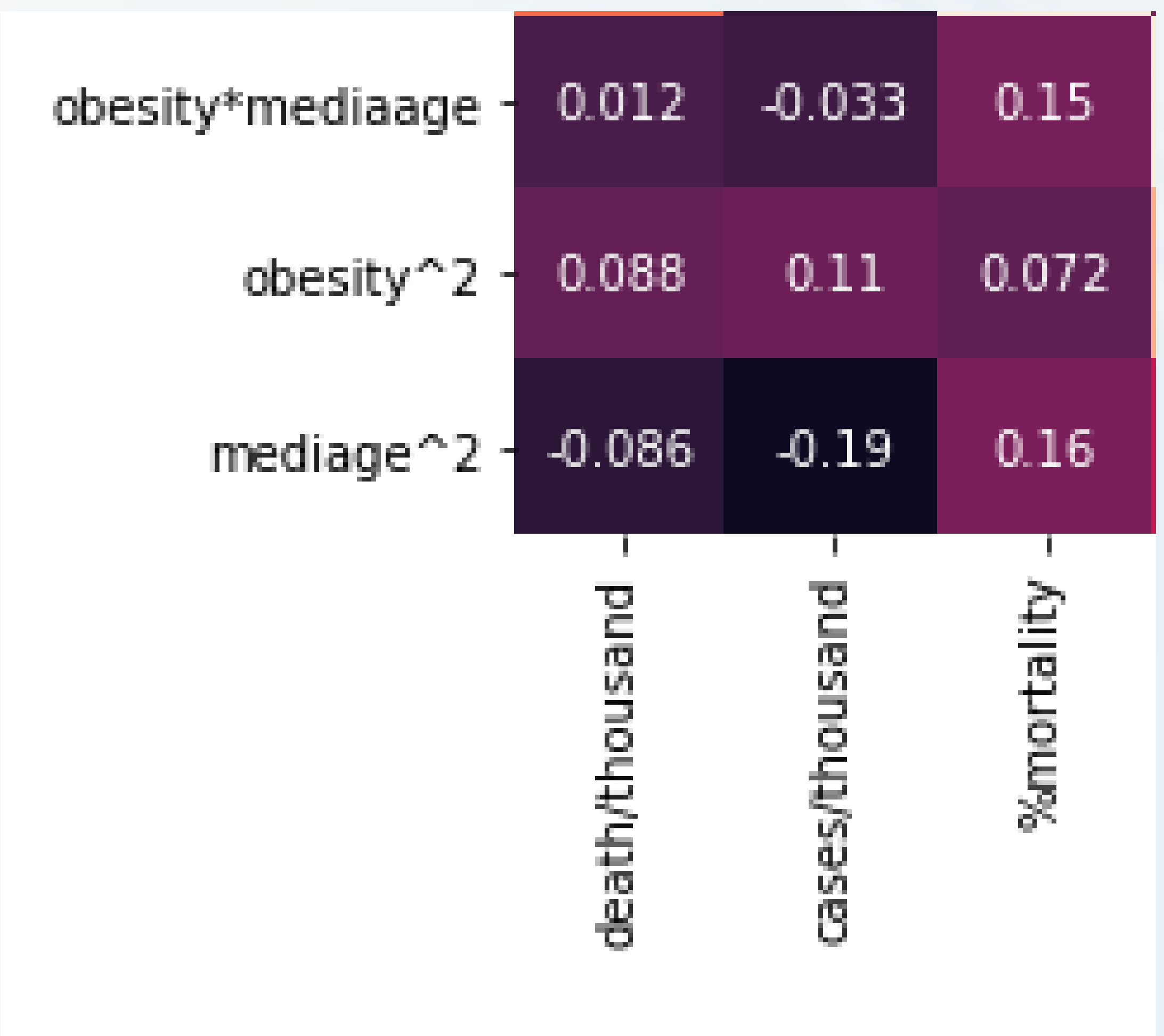
For this analysis, the latest data on the indicators of confirmed dead and recovered from covid was obtained and the obesity columns of the population and average age were added to them to relate and try to find any correlation between the indicators of covid and these demographic indicators .



	confirmed	death	recovery	population	mediage	%mortality	cases/thousand	death/thousand	Obesity
Country									
Albania	19445	559	12092	2866376	33.5	2.874775	6.783827	0.195020	21.7
Andorra	4325	75	4248	77006	36.0	1.734104	56.164455	0.973950	25.6
Austria	83267	1411	91719	8847037	43.3	1.694549	9.411852	0.159488	20.1
Belgium	333718	13055	29651	11422068	39.5	3.911986	29.216951	1.142963	22.1
Bosnia and Herzegovina	41596	1510	30939	3323929	40.1	3.630157	12.514106	0.454282	17.9

I compile them all in a single table to later evaluate the correlation in their variables, Columns = indicators, rows = European countries. Only the first 5 rows are shown here.





Observed relationships
The values that we obtain when carrying out the Pearson correlations between the obesity and mean age variables are not close to 1 or -1, this means that these variables do not influence the
% fatality, number of deaths or confirmed by country, neither in a linear nor in a quadratic or combined form

Conclusion:
We conclude that other indicators could be found that better infer the covid indicators. Below we will list them and explain how they could be used and evaluated.

Necessary data and indicators to be able to advance in a more in-depth analysis.

Temporary:

Analysis of restrictions on the movement of people by countries:

1) People traffic data (eg from data table: air traffic passengers / day by country, cars entering / leaving different countries)
2) We would compare it against the confirmed curve, analyze how the restriction measures help reduce contagions and estimate in how many days these measures take effect and what proportion of the total (eg after restricting both local and regional travel in how many days contagions decrease)
3) Then we would obtain a model to predict based on the traffic data that we would train so that it can predict the contagion curve based on parameters obtained from the traffic.
4) Also these generated graphs could be compared with the economic indicators affected by restricting travel and movements. And analyze from the economic point of view.
5) Measure the effectiveness of these models using example regression metrics (MSE, Mean Absolute Error, R) and conclude if they are feasible to apply in the future

Climate analysis:

1) Weather data (eg (temperature per day, humidity per day, ml of rain per day) per country)
2) We would compare it against the confirmed and dead curve, we will analyze if the climate is a fundamental factor, if there is a decrease due to high temperatures or high mortality in low and humid temperatures, which I would classify by zones and then see how it was in each zone (countries with similar climate parameters) the growth of the virus occurred and compare its speed.
3) If a relationship is found between these data and the covid19 indicators, regression models and classification by zones with the same climate will be tested in order to understand the expansion behavior of the virus in zones with similar climates.
4) We would apply regression metrics (MSE, Mean Absolute Error, R) and classification (Recall Confusion Matrix, F1, accuracy). If we obtain good results in the metrics, it would be used to predict new infections based on the weather.

Health system analysis

1) Health system data (eg availability of technology and hospitals, availability of medical personnel,% saturation of the health system)
2) We would compare it against the% mortality curve, and if it varies by modifying the health factors, its variation would be measured.
3) The models that we would use would be unsupervised clustering models that find some kind of relationship between these health factors and the covid indicators and then separate them and be able to study them and regress the health factors that do influence mortality.
4) Then regression models would be applied to these factors and their metrics to be able to measure the effectiveness of these models.

Timeless:

Demography of infected, recovered and dead, economic and social indices.

1) Age data of those infected and age of those killed by covid.
2) We would classify by age the infected, their mortality by age and the percentage of infected of this age,
3) With these parameters we would determine in the future the mortality% of each country based on the age of the infected.
4) We would compare this with the average mortality of each country from the data already obtained and we would determine based on which demographic group is currently being infected the probability of having a high or low mortality. These values would serve to locate where to exercise greater control to contain the virus in demographic groups with more mortality.
5) We will also correlate these values with the obesity and economic level indicators to obtain a better relationship. economic indexes for each country.

Bibliography:
Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, Author: Wes McKinney,
Recursos web:
https://covid-tracker-us.herokuapp.com
https://www.who.int/
https://data.worldbank.org/
https://www.w3schools.com/python/default.asp
Coursera :Applied Machine Learning in Python Universidad de Míchigan

Utilized Tools:
Python
Jupyter
Excel
Canvas

Python used library:
Numpy
Matplotlib
Seaborn
Sklearn
Pandas
Request