Data Science Capstone Project

Harvard University, edx

Pedro J. Llanos

Abstract

This project was part of the final Data Science Capstone course and highlights the use of data science and machine learning tools for COVID-19 applications. Due to the vast data on COVID-19 available we provide an overview on the data science worldwide, however, this current study emphasizes more on the European countries and Spain, one of the top countries leading the most number of deaths per million people due to COVID-19. The worldwide data can be obtained from: https://datascience.nih.gov/covid-19-open-access-resources, a centralized repository of the most up-to-date datasets with regards COVID-19. European data is extracted from this dataset, set from March to December of 2020. Additionally, the datasets for the Spain study was obtained from https://covid19.isciii.es/, a repository dataset (March-December 2020) that includes fifty-nine COVID-19 reports. Such reports were provided to the Red Nacional de Vigilancia Epidemiológica (RENAVE) by the National Centre of Epidemiology (Instituto de Salud Carlos III- Centro Nacional de Epidemiología). These reports were processed with Excel, then read as "csv" files by the R-algorithm. This report shows preliminary machine learning applications to several scenarios (Europe and Spain) while providing a comparative analysis between different models for each scenario.

I. INTRODUCTION

By the end of 2020, the world was struck by one of the deadliest viruses yielding almost 100 million cases and 2 million deaths due to COVID-19. This is just the beginning of an evolving virus that will still expand its footprint worldwide in several years to come before it is finally "controlled". The motivation behind this project was born early in 2020 when the pandemic was confirmed and very few people saw it coming while the rest of the world was being oblivious of the subsequent consequences of this virus.

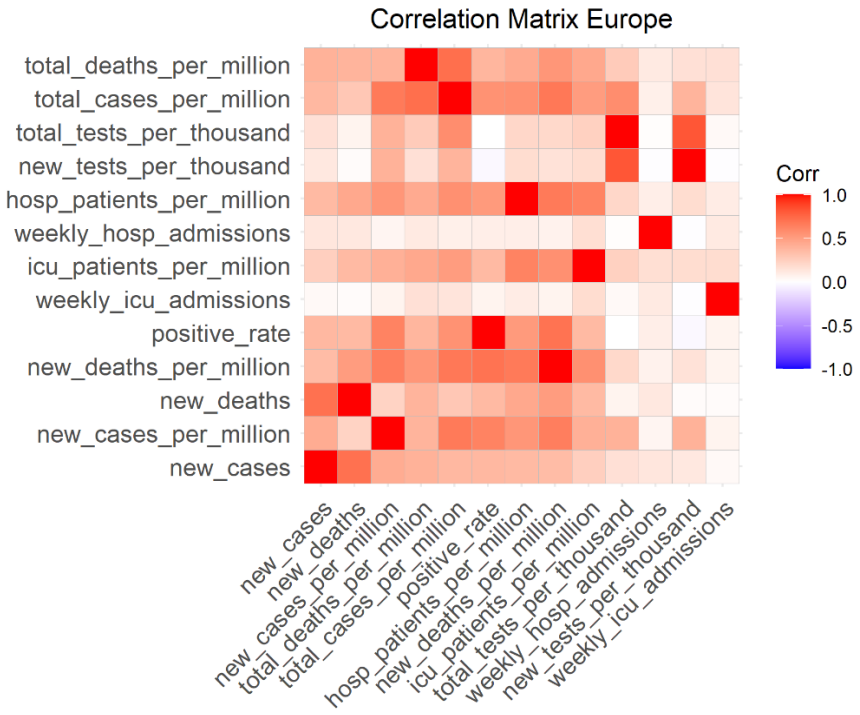
In the first part of this project we conducted a preliminary analysis of the trends of COVID_19 worldwide. The second part of the project is dedicated to further analyze these trends for Europe with some machine learning application. The third part of the project is tailored towards a more specific case in Europe, Spain, a leading country in the number of deaths per million with also some machine learning applications. Spain is a country with 19 autonomous communities (CCAA), and we will analyze the spread of COVID-19 across each CCAA and learn about some of the most prominent trends for those CCAA leading the spread in Spain.

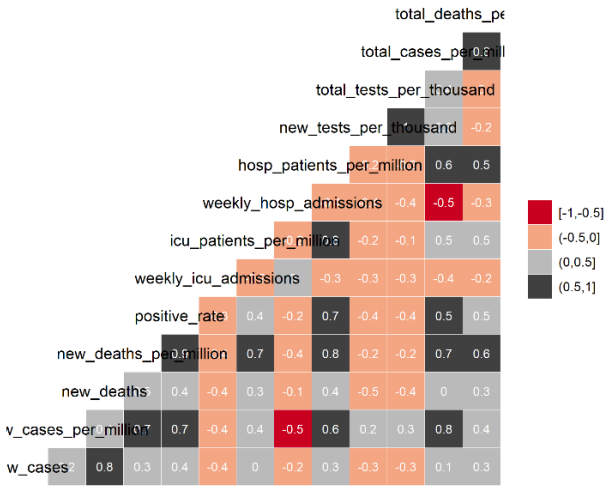
At first glance, we looked at the correlation matrix for the European and Spain studies. The correlation matrix tells us if there is any strong relationship between variables, and the

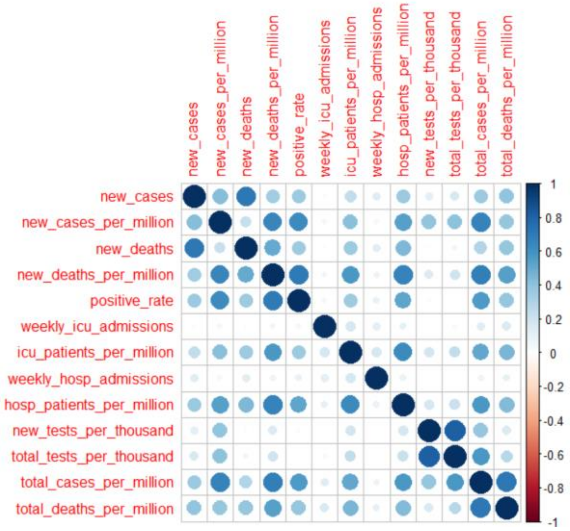
direction (positive or negative) of such relationship. Preliminary understanding of theses variables is important before applying machine learning to this large data science problem.

A. Europe study:

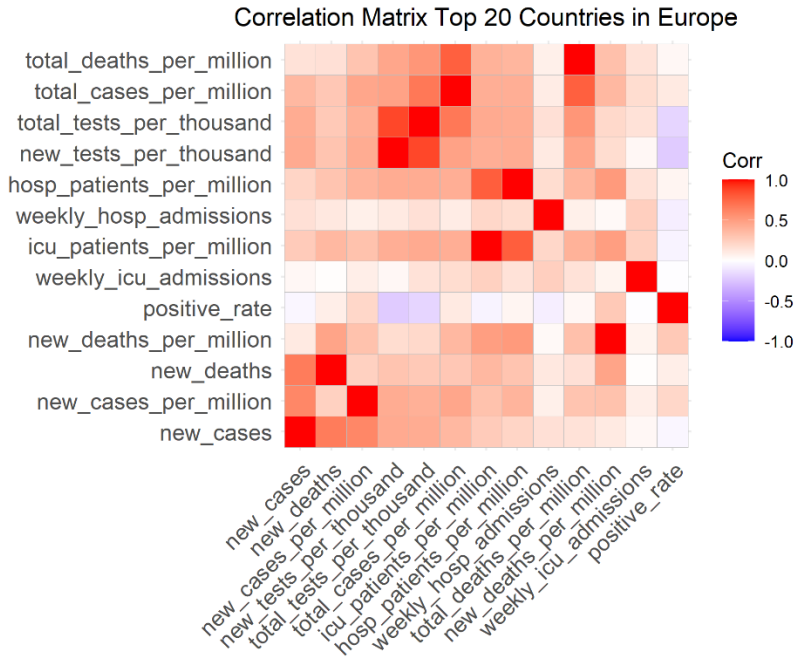
Correlation matrix for all European countries:

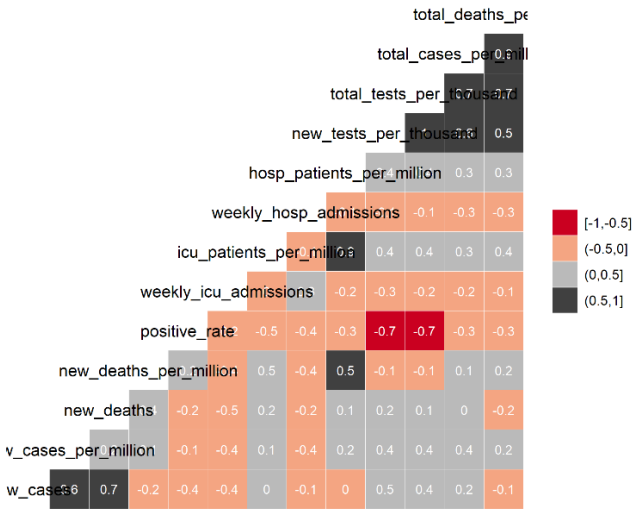


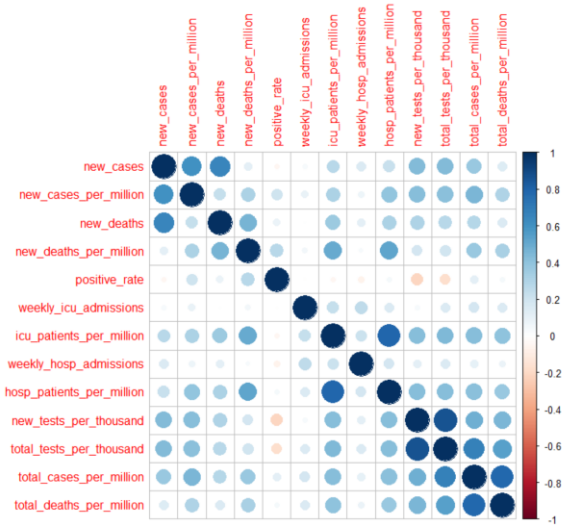




Correlation matrix for top 20 European countries:







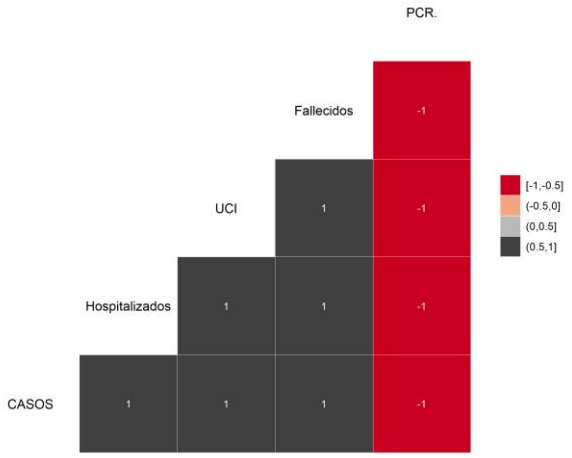
Note both correlation matrices for all European countries and the top 20 European countries can yield a quite different behavior between variables.

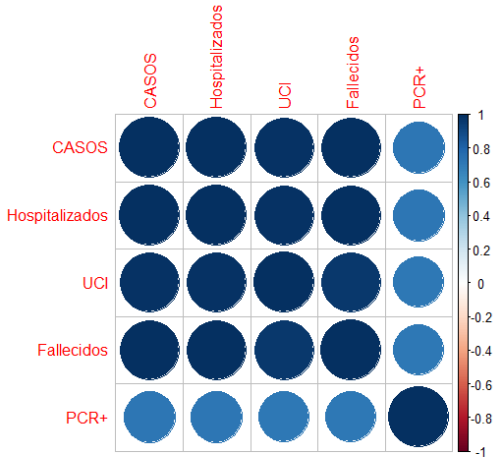
B. Spain study

Correlation matrix for all CCAA in Spain:

- CASOS = number of cases
- Hospitalizados = number of hospitalized
- UCI = number of patients in Intensive Care Unit
- Fallecidos = number of deaths

In this study we will keep the notation as provided by CNE of Spain.





Note the strong relationship between the variables.

II. METHODOLOGY

Pertinent methodologies are provided for the several analyses corresponding to the Europe and Spain studies. Europe data was filtered from the worldwide data "**covid-19-world-cases-deaths-testing.csv**" and **"LatLonEUcountries.csv".** For the Spain study, the general data was obtained from the file **"Coronavirus_Symptoms_Spain_Normalized_AgeGroups.csv"** and specific data for each CCAA was obtained from the file **"CoronavirusSPAIN.csv".**

a) Spain Study:

For the Spain study, different models are used to analyze the effect various COVID-19 variables, such as the total number of cases (CASOS), the number of hospitalized (hospitalizados), the total number in the intensive care unit (Unidad de Cuidados Intensivos or UCI), the total number of deaths (Fallecidos), and the date (FECHA). These models are described below:

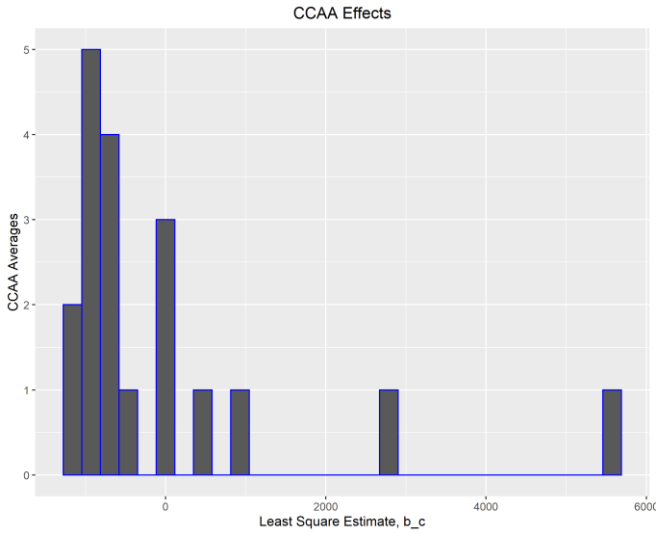
1. **First model**: Analyze the naïve RMSE with just the "Average". This model assumes that the Fallecidos ("deaths") is the same for all CCAA and hospitalized, and their differences are associated with random variations:

$$Y_{u,c} = \mu + \varepsilon_{u,c}$$

where μ is the average ("true" Fallecidos) for all CCAA, and $\varepsilon_{u,c}$ are the independent errors sampled from the same distribution centered at zero.

2. **Second model**: Analyze the "CCAA" effect. This model assumes another extra term, b_c, or CCAA-specific effect with respect to the previous model, as seen below:

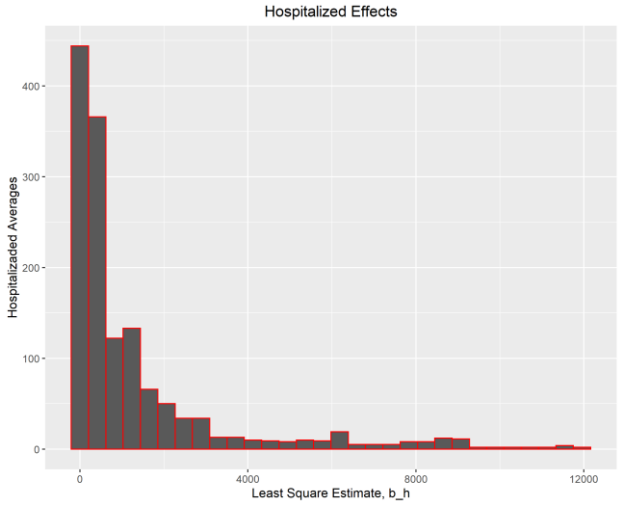
$$Y_{u,c} = \mu + b_c + \varepsilon_{u,c}$$



3. **Third model**: Analyze the "Hospitalizados" effect. This model assumes another extra term with respect the previous model:

$$Y_{u,c} = \mu + b_c + b_h + \varepsilon_{u,c}$$

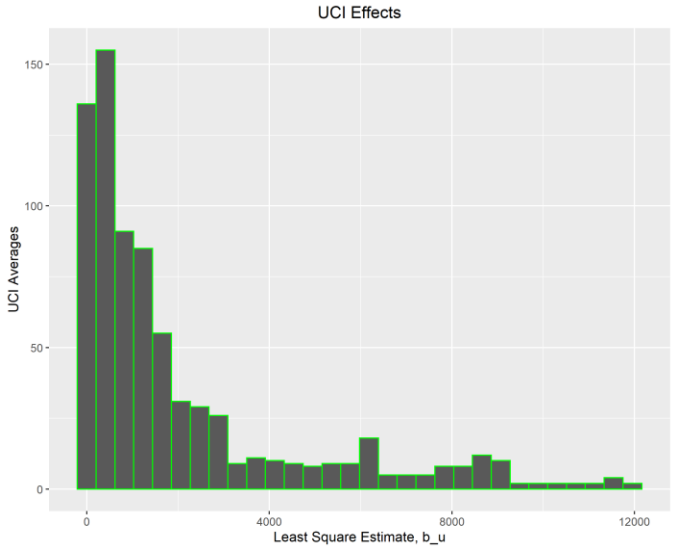
where b_h is the hospitalized-specific effect.



4. **Fourth model**: Analyze the “CCAA”, “Hospitalizados” and “UCI” effects. This model assumes another extra term with respect the previous model:

$$Y_{u,c} = \mu + b_c + b_h + b_u + \varepsilon_{u,c}$$

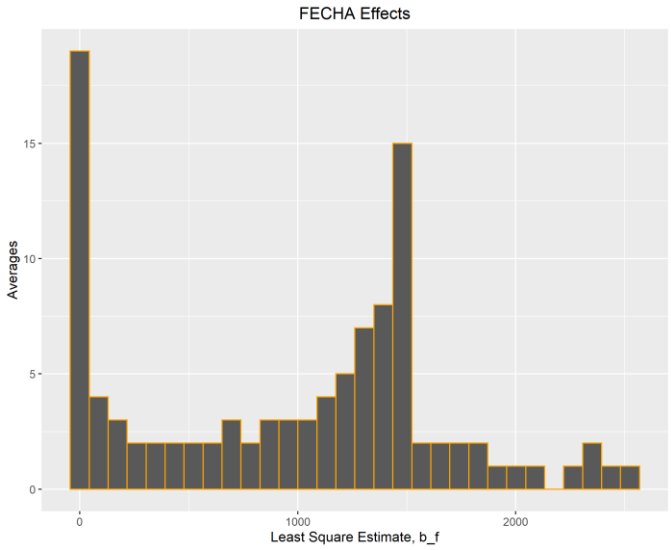
where b_u is the UCI-specific effect.



5. **Fifth model**: Analyze the “CCAA”, “Hospitalizados”, “UCI”, and “FECHA” effects. This model assumes another extra term with respect the previous model:

$$Y_{u,i} = \mu + b_i + b_u + b_g + b_t + \varepsilon_{u,i}$$

where b_t is the time-specific effect.



6. **Sixth model**: Analyze the regularization on the "CCAA" and "Hospitalizados" effects. This model is based on the minimization of the equation with a penalty, instead of the minimization of the least squares method:

$$\frac{1}{N}\sum_{u,c}\left(y_{u,c}-\mu-b_c\right)^2+\lambda\sum_{c}{b_c}^2$$

where the first term is the well-known least squares method and the second term is the added penalty. The values of b_c that minimize the above equation can be obtained as follows:

$$\widehat{b_\iota}(\lambda)=\frac{1}{\lambda+n_c}\sum_{u,c}^{n_c}\left(Y_{u,c}-\hat{\mu}\right)$$

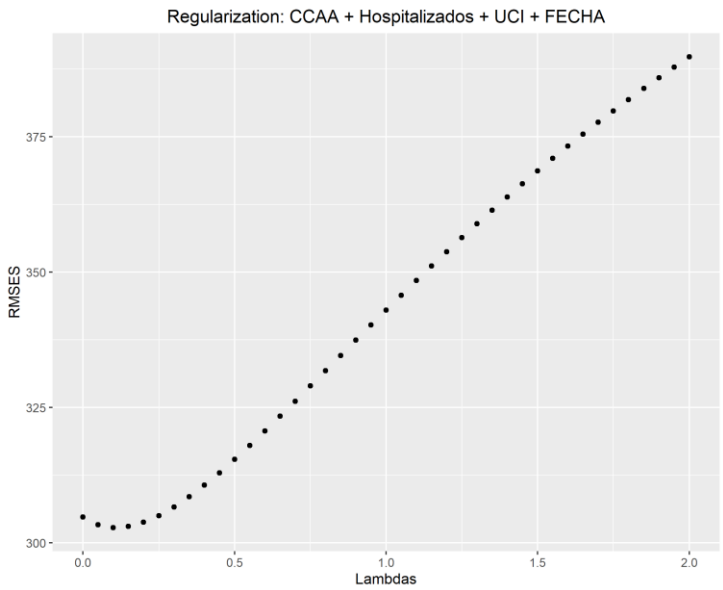
Next, we can choose the penalty terms, since we know λ a tuning parameter. We can use cross-validation to choose it. In this case we pick the movie effect for the regularization case.

Can we find another λ that minimizes even more the RMSE? For this, we can use the regularization to estimate the user effects too.

$$\frac{1}{N}\sum_{h,c}\left(y_{h,c}-\lambda-b_c-b_h\right)^2+\lambda\left(\sum_{c}{b_c}^2+\sum_{h}{b_h}^2\right)$$

$$\widehat{b_h}(\lambda)=\frac{1}{\lambda+n_c}\sum_{h,c}^{n_c}\left(Y_{h,c}-\hat{\mu}-\widehat{b_\iota}\right)$$

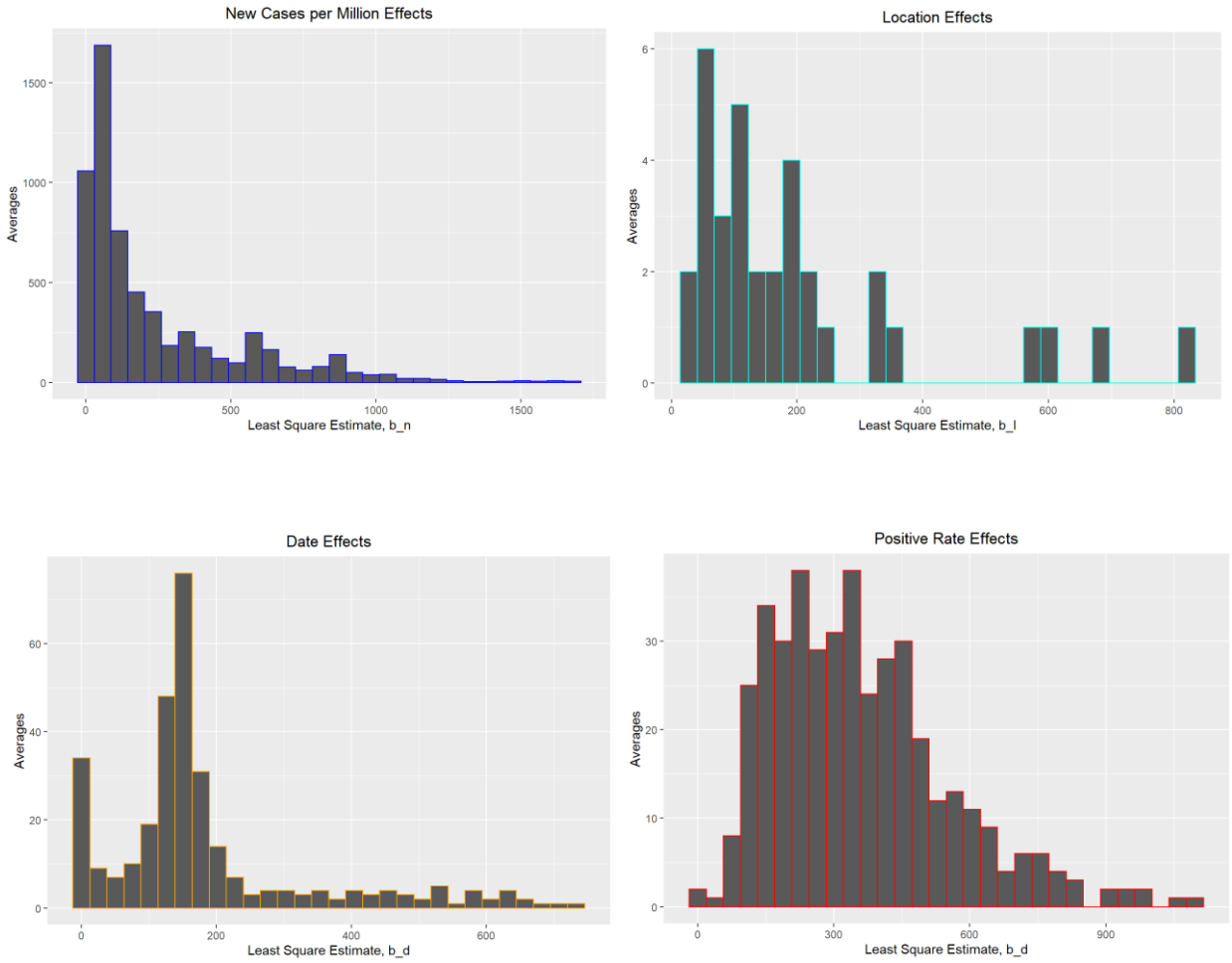
With this improved model, we get a λ that minimizes the RMSE ($\lambda=0.1$).



b) Europe Study:

Similarly, as we did for the Spain study, we used the same machine learning mathematical approach to analyze the Europe case but using different variables for each respective model. Some of the variables considered where the new cases per million, location effects, date effects and positive rate effects in order to estimate the number of deaths per million.

Regularization Europe Study:



When using regularization, the lambda that minimizes the RMSE for each case is:

$\lambda = 1.25$ (date)

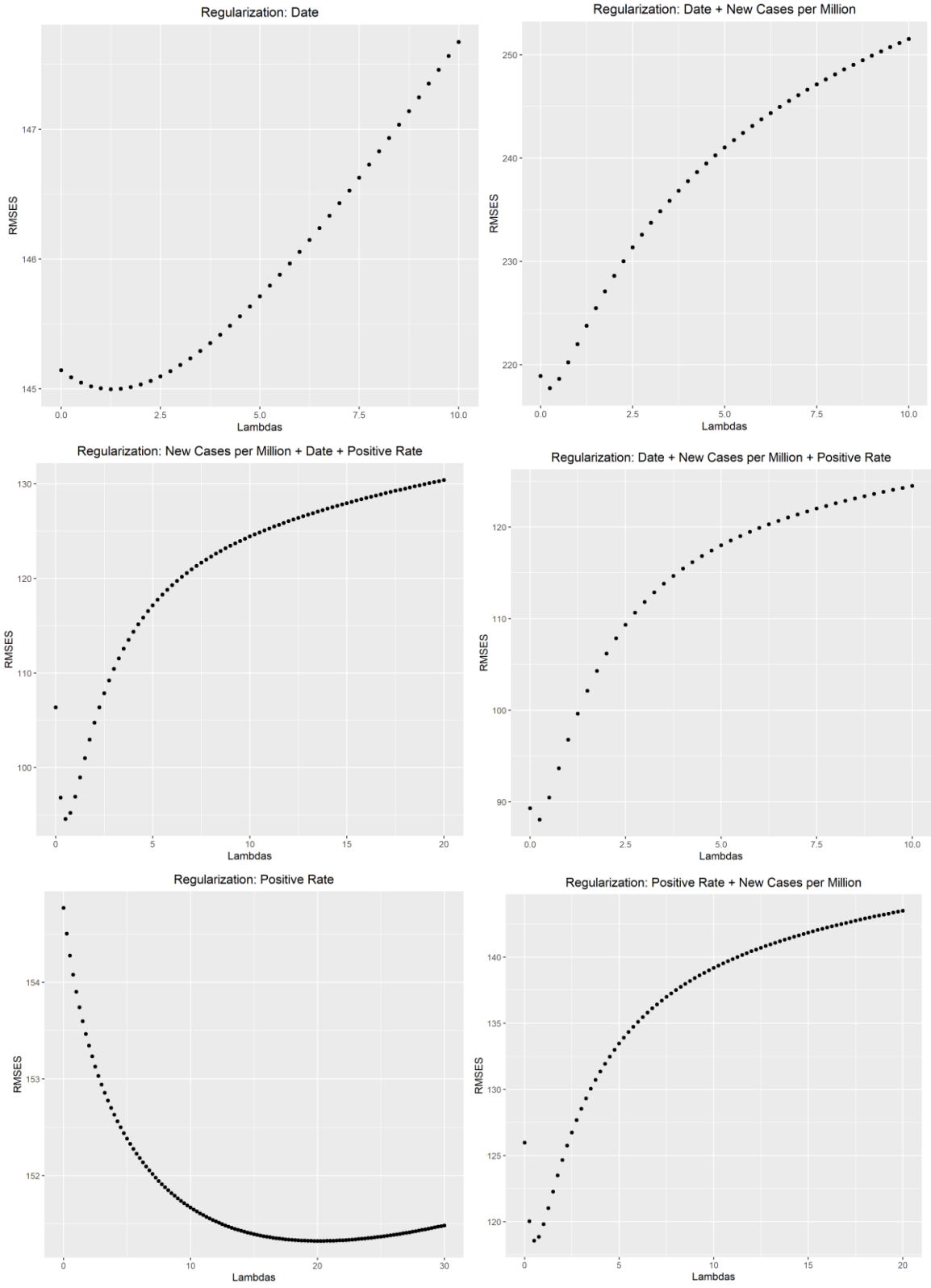
$\lambda = 0.25$ (date + new cases per million)

$\lambda = 0.25$ (date + new cases per million + positive rate)

$\lambda = 0.5$ (new cases per million + date + positive rate)

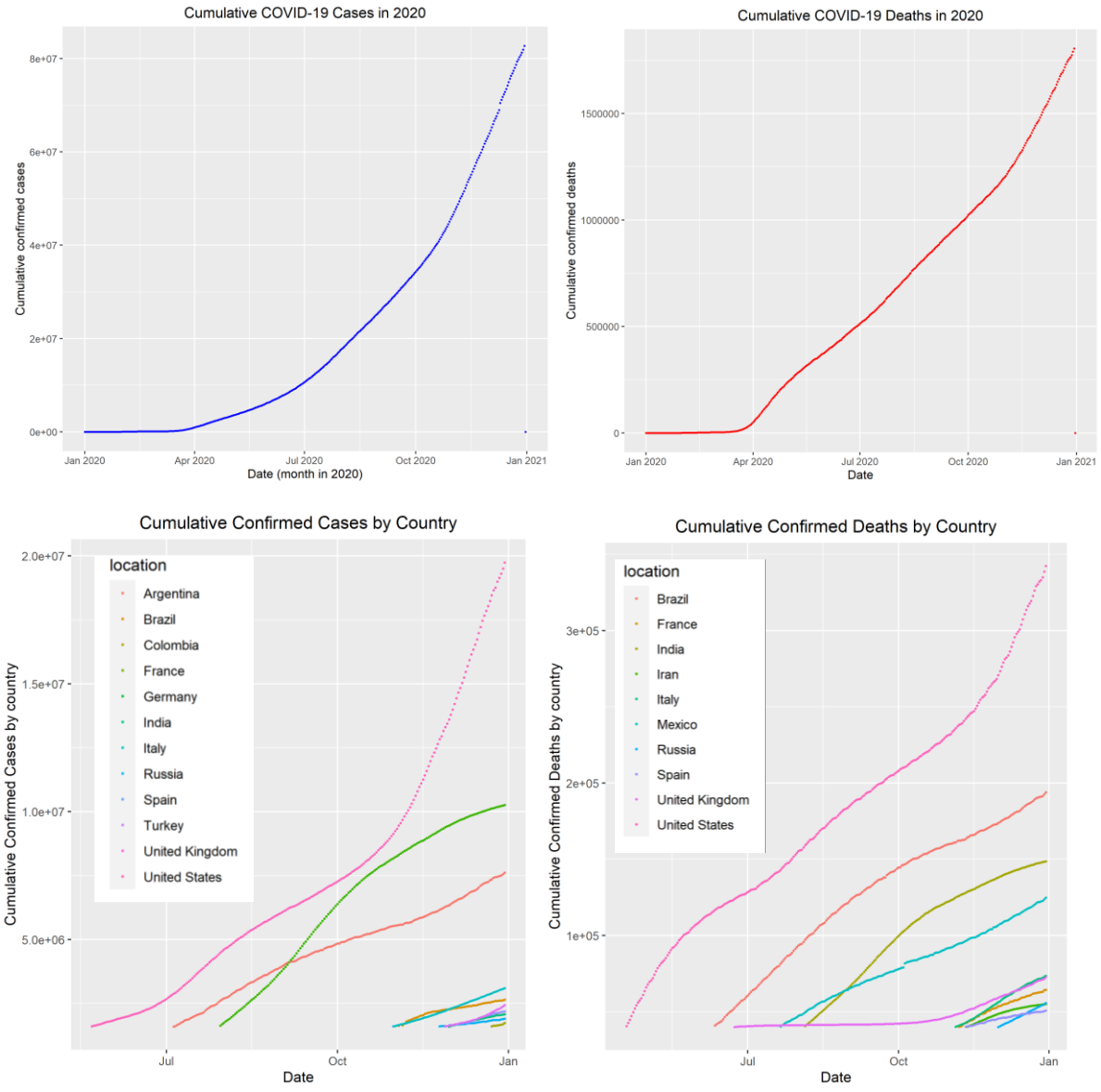
$\lambda = 20.25$ (positive rate)

$\lambda = 0.5$ (positive rate + new cases per million)



III. RESULTS

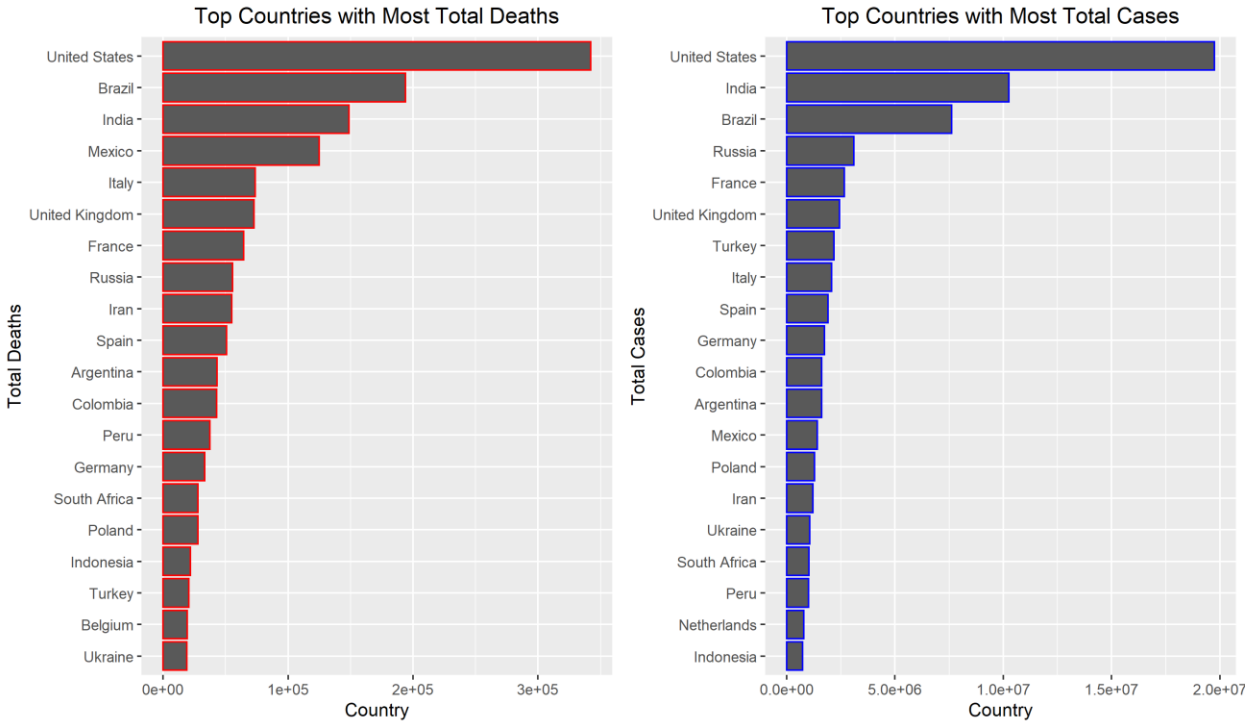
During the preliminary inspection of the data, we looked at the top 20 countries with highest total number of cases and deaths shown below:

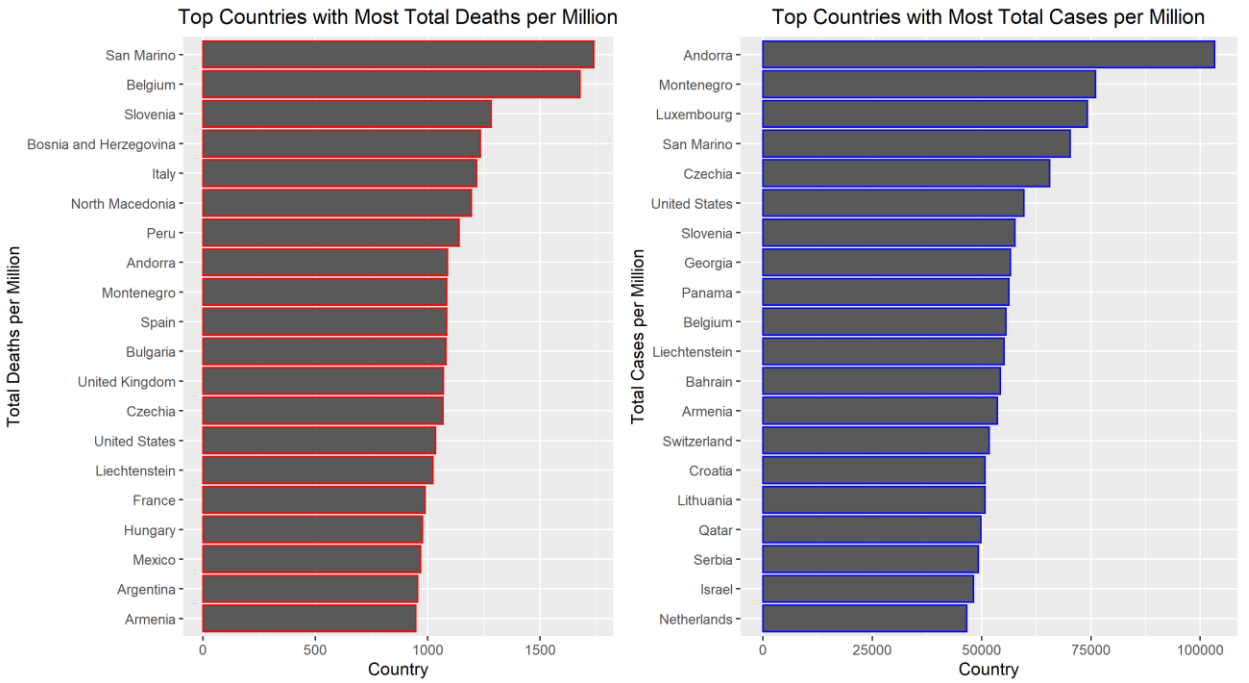


How is Spain (about 1.3 times smaller than Texas) doing with respect the rest of the world? For this, we will look at the total cases per million and total deaths per million.

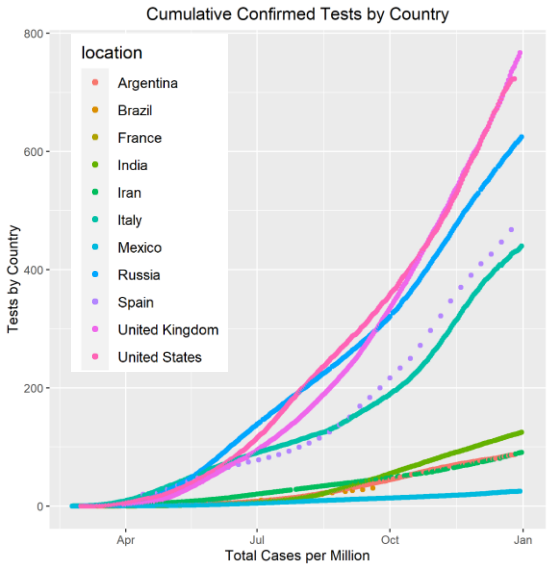
Although Spain is among the top 10 countries leading in total cases and deaths worldwide, it does not appear among the top 20 countries with the most total cases per million. However, Spain is one of the top 10 countries leading with most total deaths per million. Other countries, such Luxembourg have very large number of cases per million, yet they do not appear as top 20

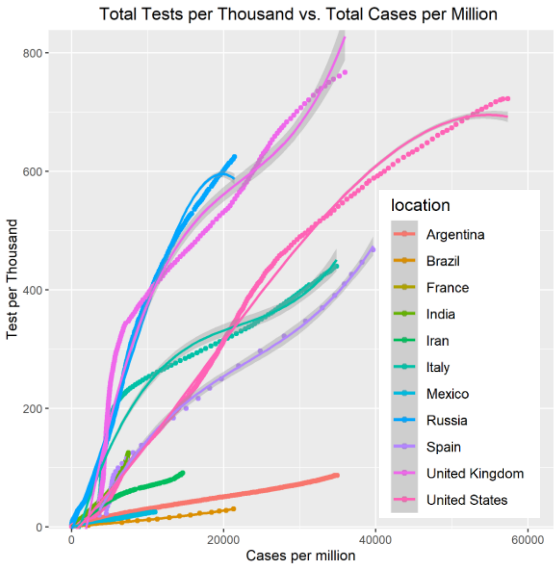
countries with highest number of deaths per million. Luxembourg is one of the countries with highest number of tests.



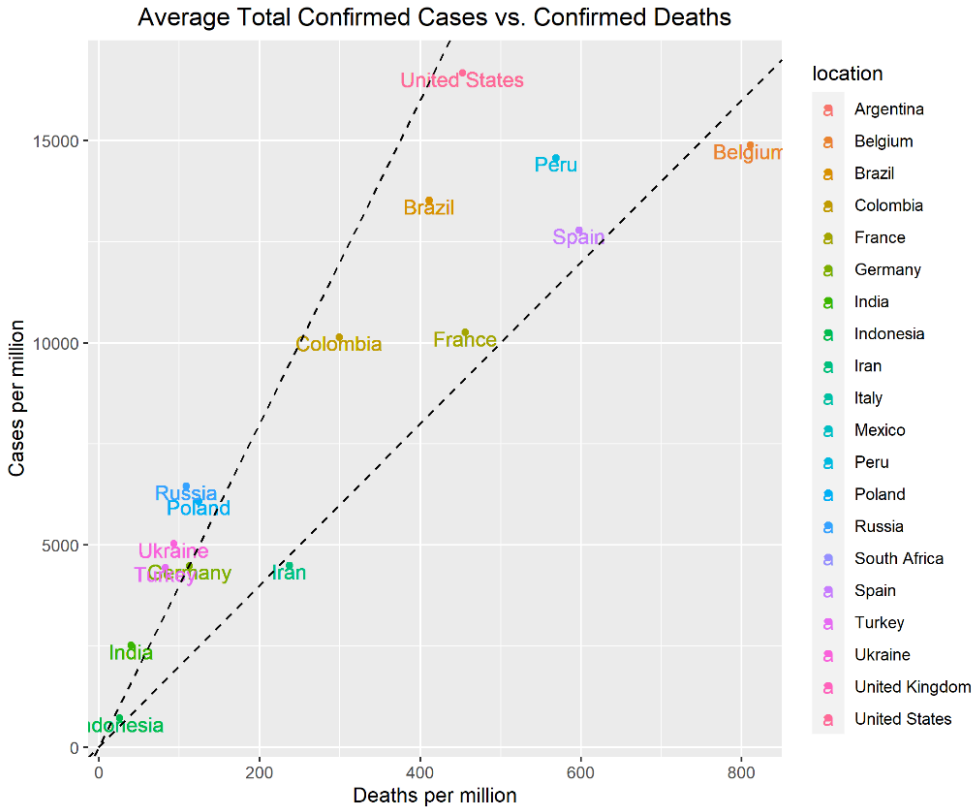


Spain is also one of the top 4 countries with more tests conducted and with tests per thousand versus the total cases per million. Spain has one of the steepest slopes.

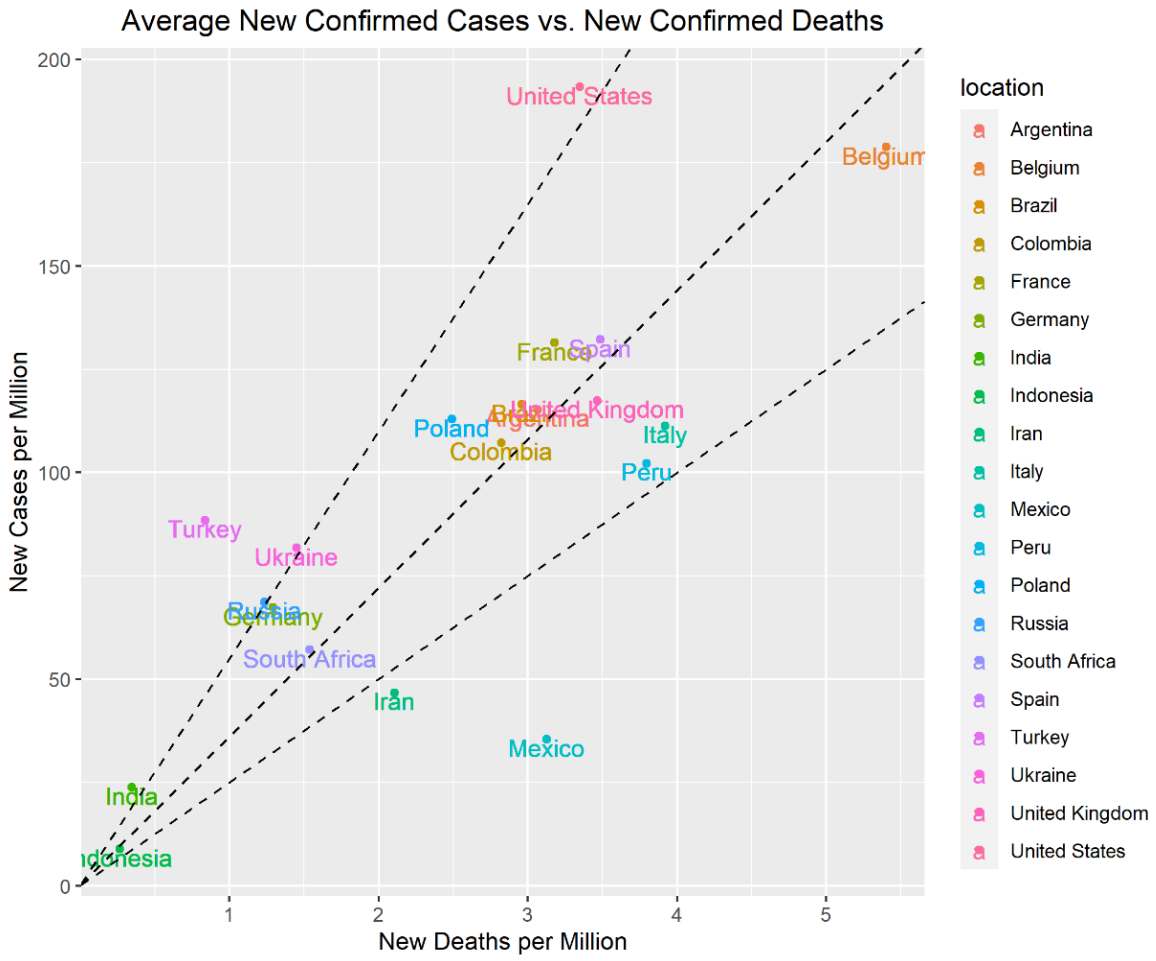




The next plot shows the average confirmed cases versus the confirmed deaths for the top 20 countries. On the plot, two imaginary trend lines are depicted showing at least 10 countries following the steepest dashed line and the rest of these top countries following the less steep dashed line. Top right corner of the plot shows European countries (Belgium, Spain, Italy, France, UK) with the highest average of number of deaths and cases per million whereas the bottom left corner of the plot shows most of the top countries with the lowest average of confirmed deaths and cases per million.

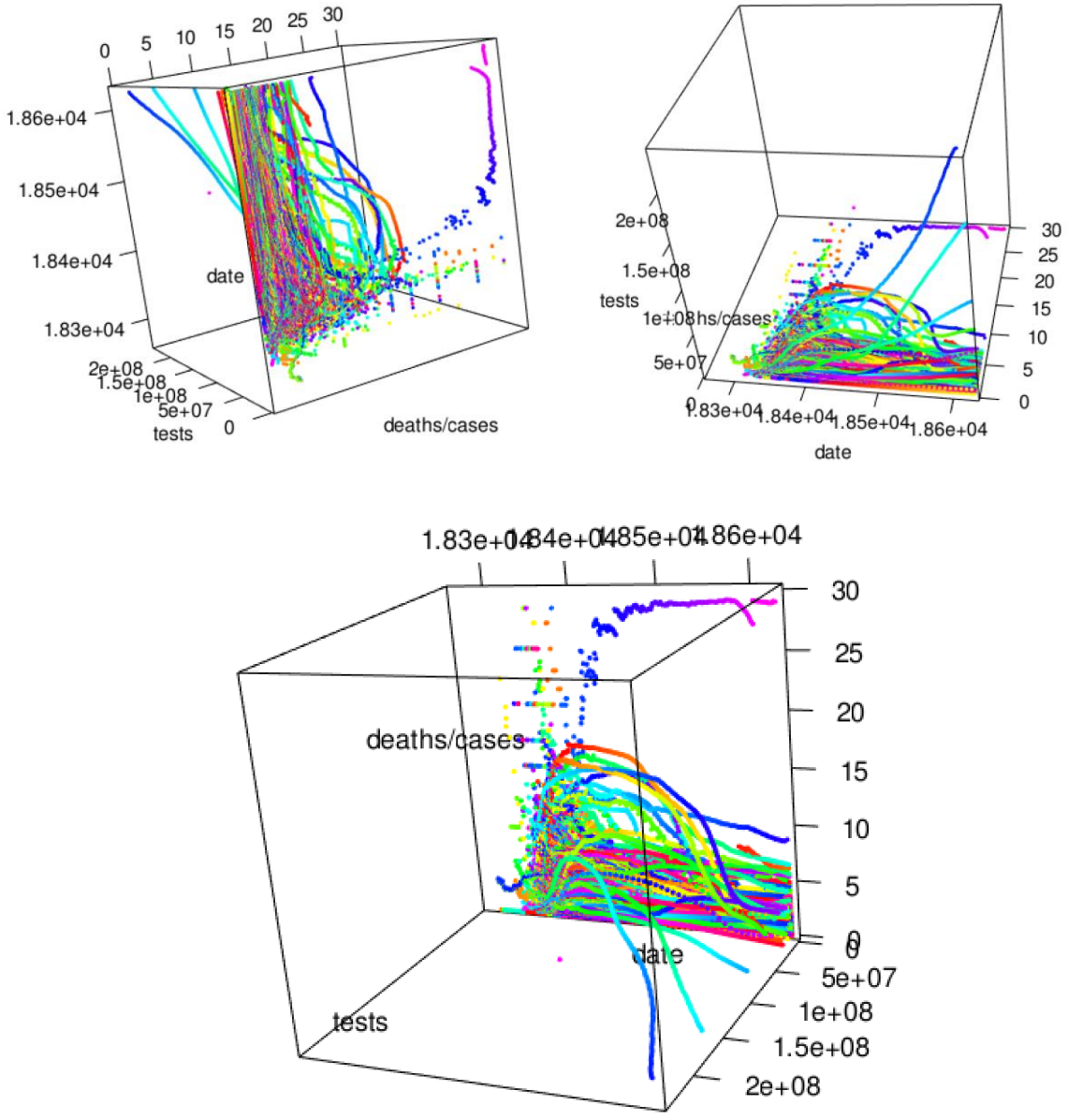


The next plot displays a similar analysis but considering new confirmed cases and new confirmed deaths for the top 20 countries. Three preliminary dashed lines were drawn on the plot to indicate possible trends of some countries. We see that some of the top countries with highest number of cases and deaths per million show the highest number of new cases and deaths per million. Most of these countries follow the middle-dashed line trend while fewer countries follow the upper or lower dashed lines.



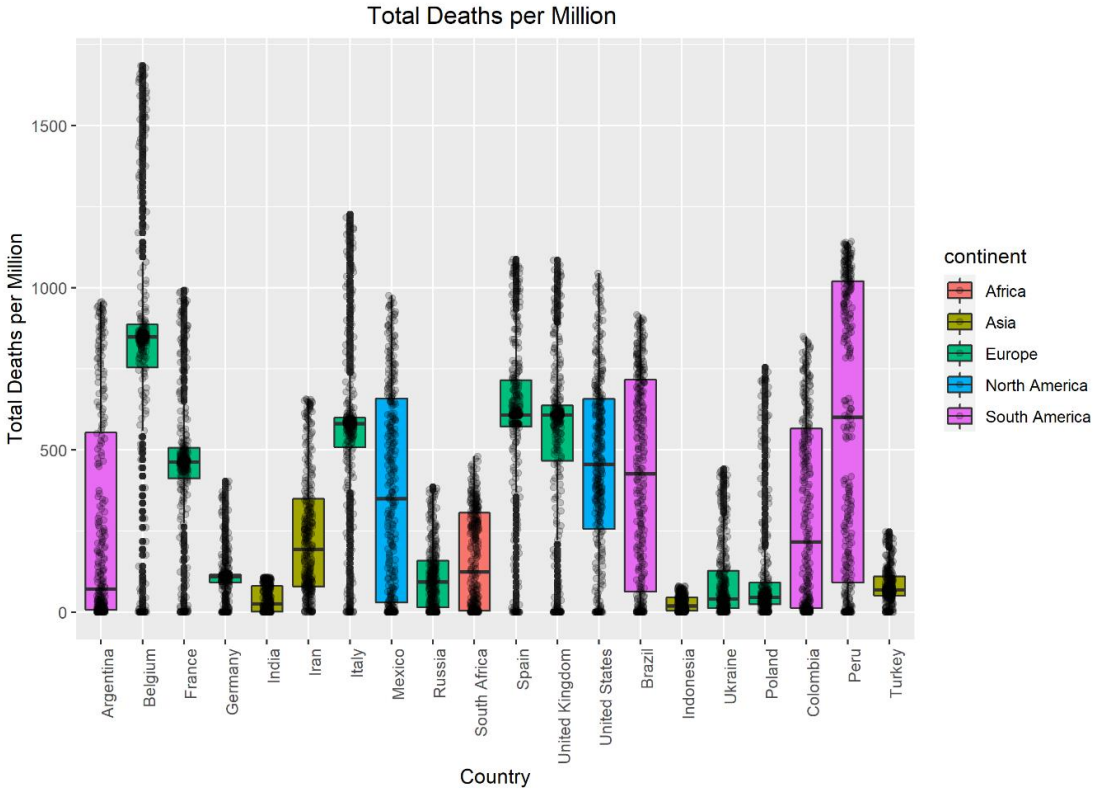
As countries evolve in time, the relationship between new cases per million and new deaths per million tend to be move towards one dashed-line trend or another, being the middle one the one that most countries tend to converge towards.

Although the 3D graph may not be very useful to extract any direct clear observations, it may be helpful at times for smaller datasets for epidemiology purposes. This is not further pursued in this report.

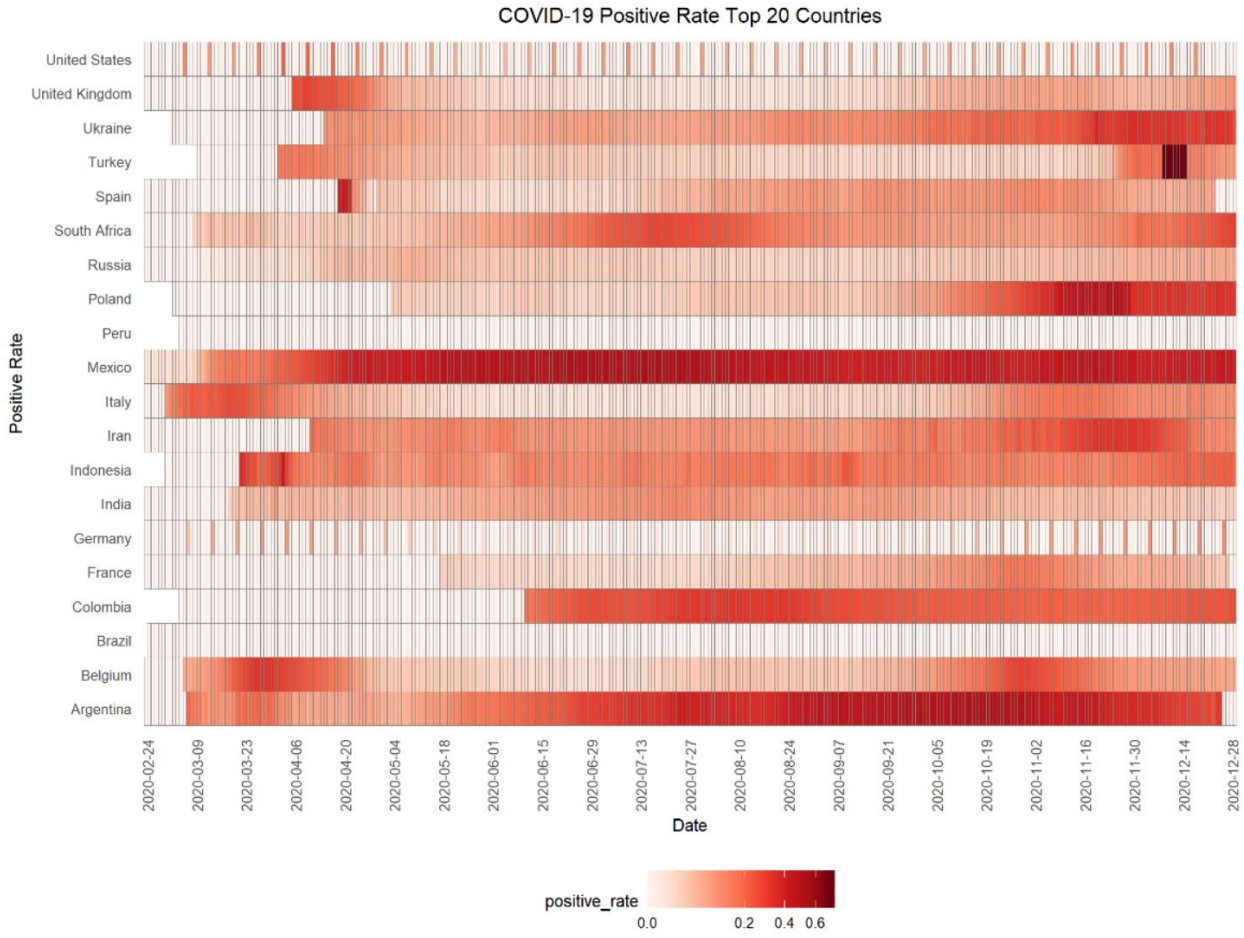


From the next graph, we can extract some observations from the top 20 leading countries in deaths per million in the world. The horizontal line represents the median, which divides the lowest 50% of observations from the upper 50% of observations. The bottom part of each of the boxes below represents the first quartile (Q1), so below Q1 you have the lowest 25% of observations. The top part of the boxes represents the third quartile (Q3) or the upper 25% of observations from the given data. Some observations are:

a. European countries have higher median and longer spread of lower and upper values generally.
b. European countries have smaller boxes (more condense boxes) than other countries from other continents, so the variation is less.
c. North America and South America countries have larger boxes or much larger variation of total deaths per million.
d. Most European countries 25% of their observations comprise all the observations of most Asian countries.
e. Belgium lowest 25% of observations comprises the upper 25% of observations from the United States.
f. Q3 for United Kingdom and the United States are almost identical, yet the mean of deaths per million for United Kingdom is about 30% higher than for the United States.
g. Lowest 50% of the observations for Belgium represent almost all the data observations for the North America and South America countries.

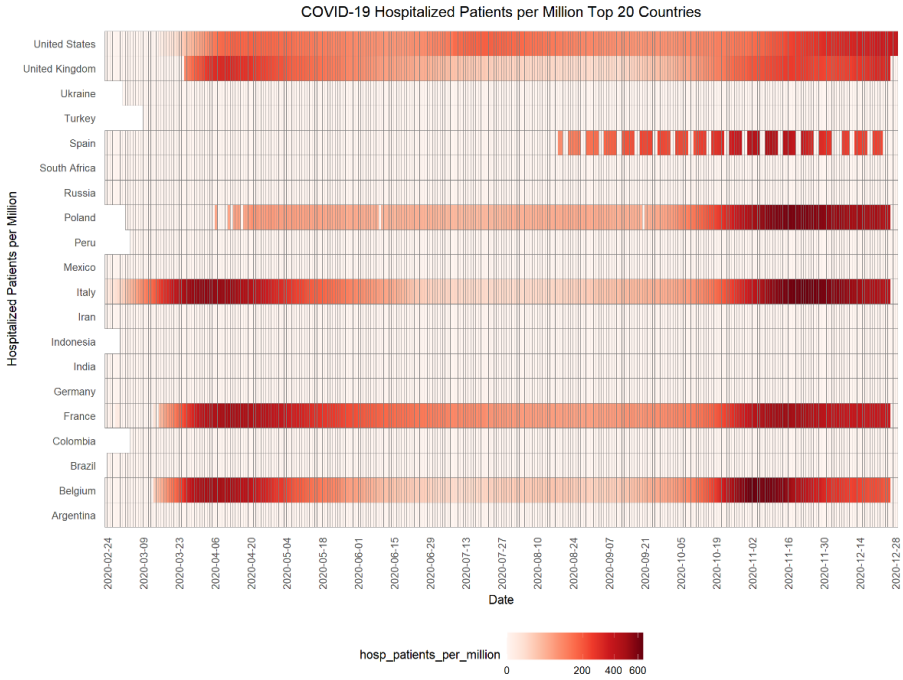


Next, we show some tile plots for the top 20 countries with highest positive rate index, highest number of hospitalized patients per million and largest number of new cases per million:

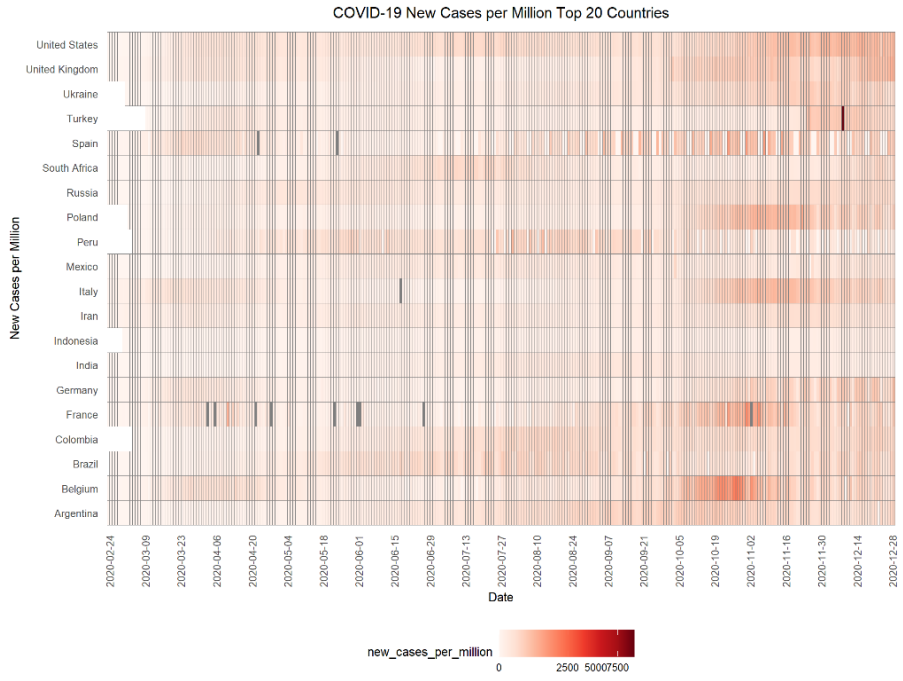


The positive rate (or ratio between the positive tests and total tests x 100%) gives an idea of the actual level of coronavirus transmission in the country and whether that country is conducting enough tests for the amount of people for that specific country. For example, Spain, being one of the countries with most deaths per million has not conducted enough tests as compared with other countries.

Similarly, the tile plot showing the number of hospitalizations per million for the top 20 countries. The United States, United Kingdom, Italy, Spain, Poland, France and Belgium appear to be the top countries with most hospitalized patients among these top 20 countries. Note the intermittent behavior of hospitalized patients per million during the last four months in Spain while other countries show a steadier rate of hospitalizations per million. Another observation is that during March-May there was a higher rate of hospitalized patients per million, then these number receded from June-September, and climbed back again in October-December of 2020.



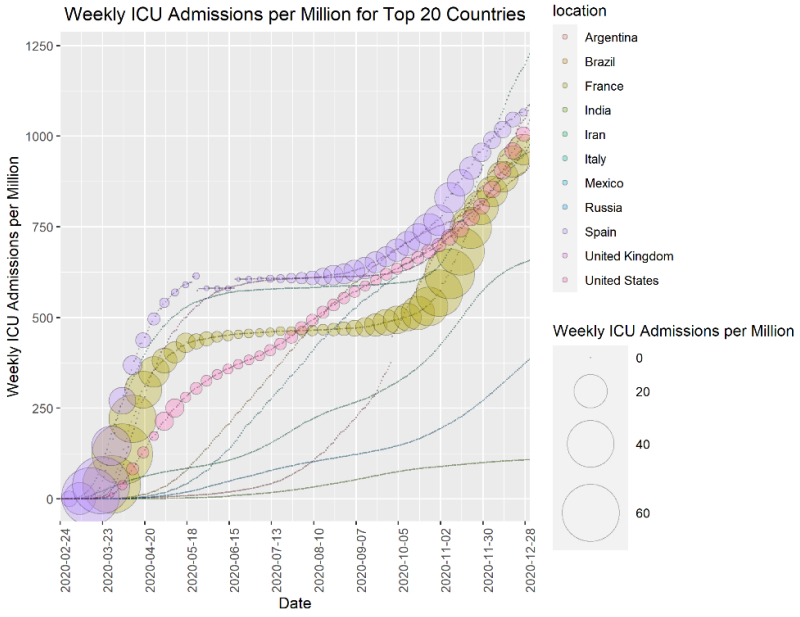
Similarly, we can visualize the new cases per million for the top 20 countries with most cases.

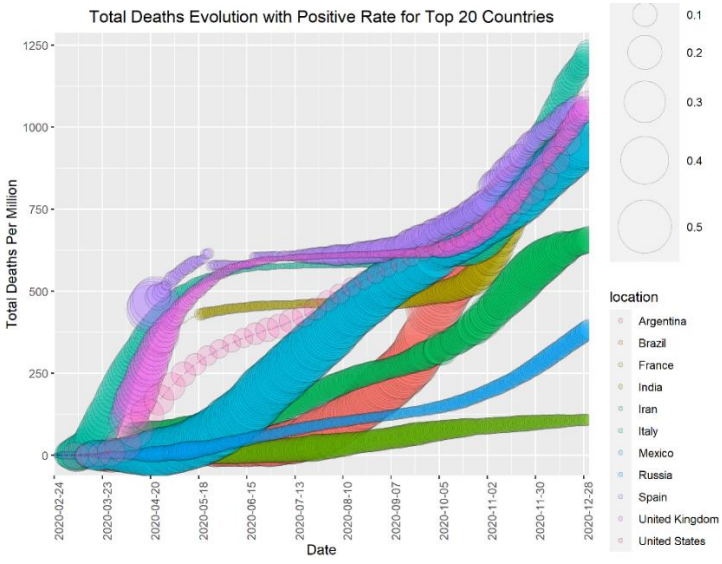


The United States shows a quite steady rise of new cases per million since November of 2020. It also shows several rise pulses of new cases per million in April, July and November-December of 2020. Other countries show a similar behavior having some first waves of new cases per million around March-April, then around varying months in the summer, and most countries show another rise wave of new cases per million between October and December 2020.

Some variables can sometimes be better visualized with other type of plots. Thus, we are using bubble plots here to visualize the ICU patients for top 10 countries. Some visual observations can be obtained from this plot:

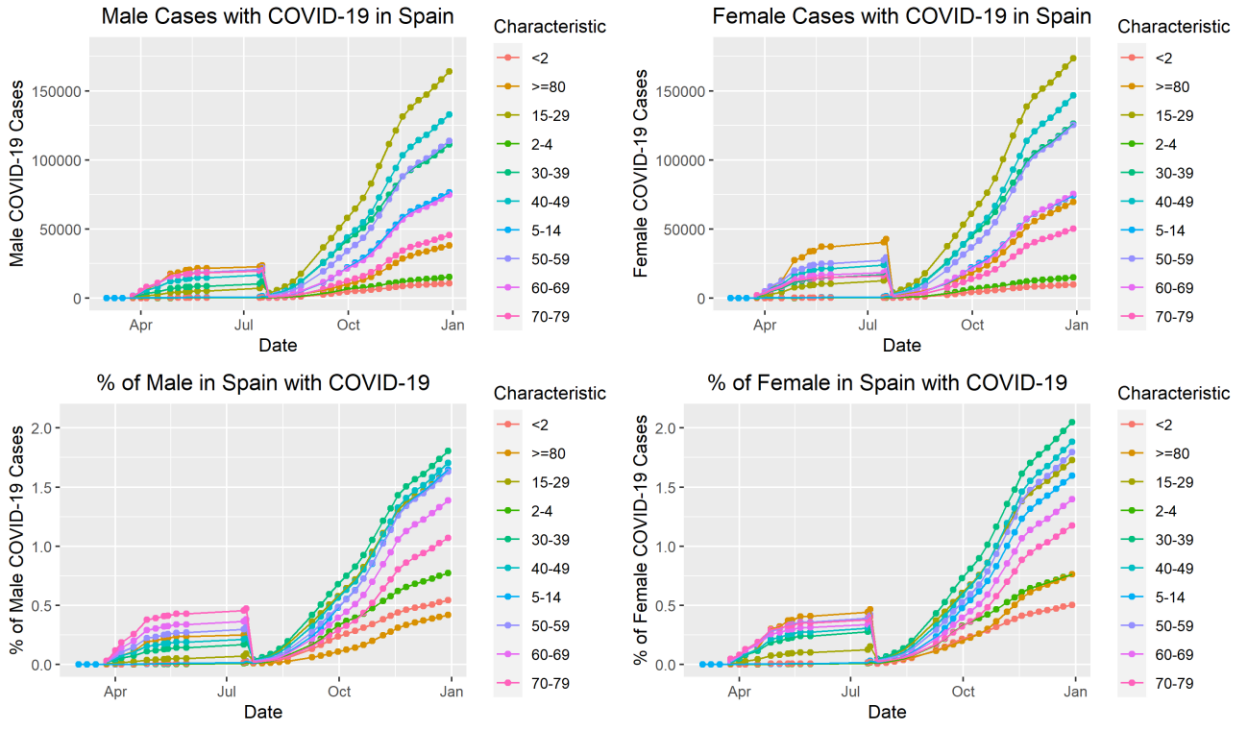
a. France and Spain have the largest weekly ICU admissions per million. Note these bubbles were the smallest around end of May-June timeframe but started to increase back again toward the end of the summer after reaching an inflexion time period of about two months.
b. The United States shows a steadier number of weekly ICU patients per million, it almost grows linearly.
c. Spain number of deaths receded back in May-June timeframe, yet it was the country with highest number of deaths per million. To date, Spain is still a leading country in the number of deaths per million.





Let's analyze some results for Spain. Data obtained for Spain came from 59 different reports in 2020:

- These reports were very insightful, but they did not have the same structure or were organized in the same manner that previous reports were.
- At least 5 types of different style reports have been written for all reports what made it very difficult to automate things in a timely manner for this study.
- Many reports have some numerical inconsistencies, yet insignificant for statistical purposes.
- Initial reports reported different age groups than later reports, so we do not know if "65+" age group is for 60-69 years old, 70-79 years old or >= 80 years old. This is also applicable for other age groups too. Thus, for this reason we decided not to include this data. This data may represent only less than 4% of all the observations.
- Some reports stopped reporting very relevant data, such as the different symptoms associated with COVID-19 encountered for the Spanish population.

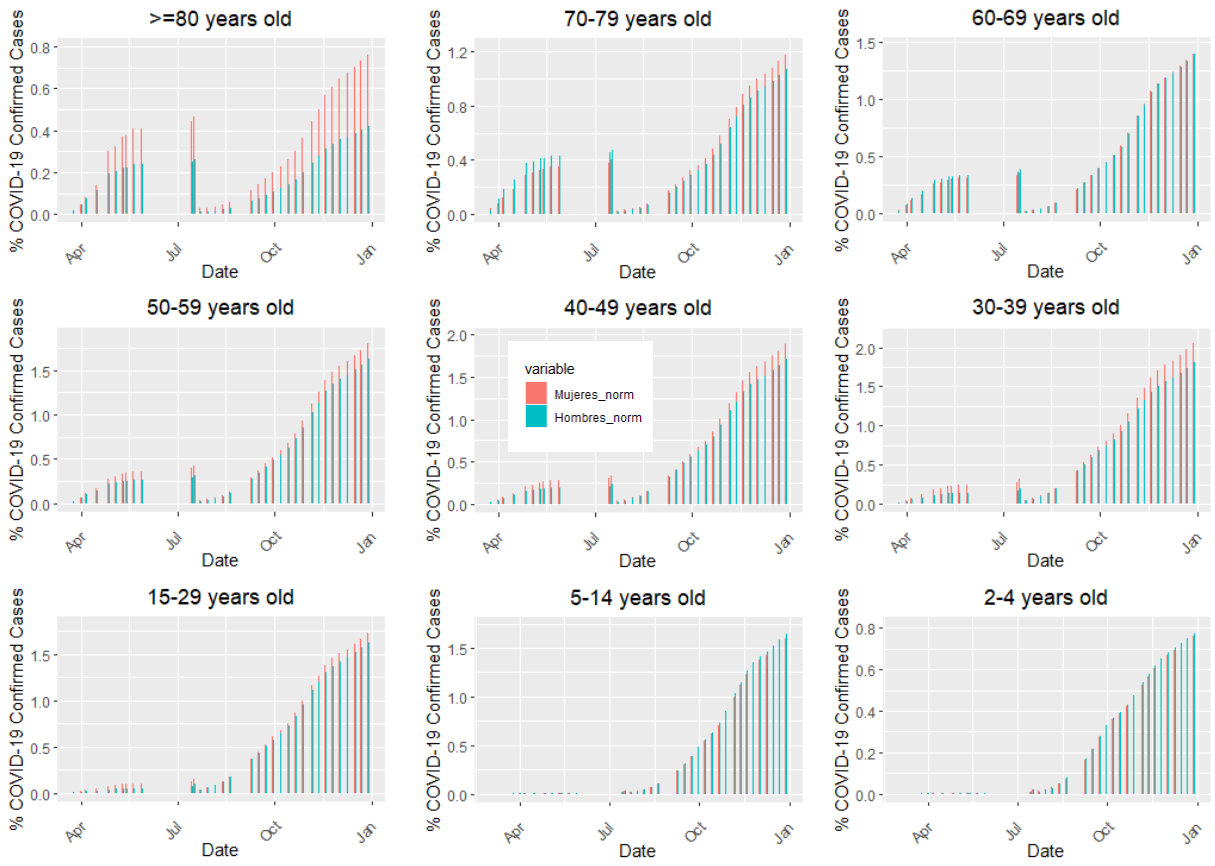


The previous graphs have been normalized with age groups by considering the population pyramid of Spain obtained in https://www.populationpyramid.net/spain/2019/.

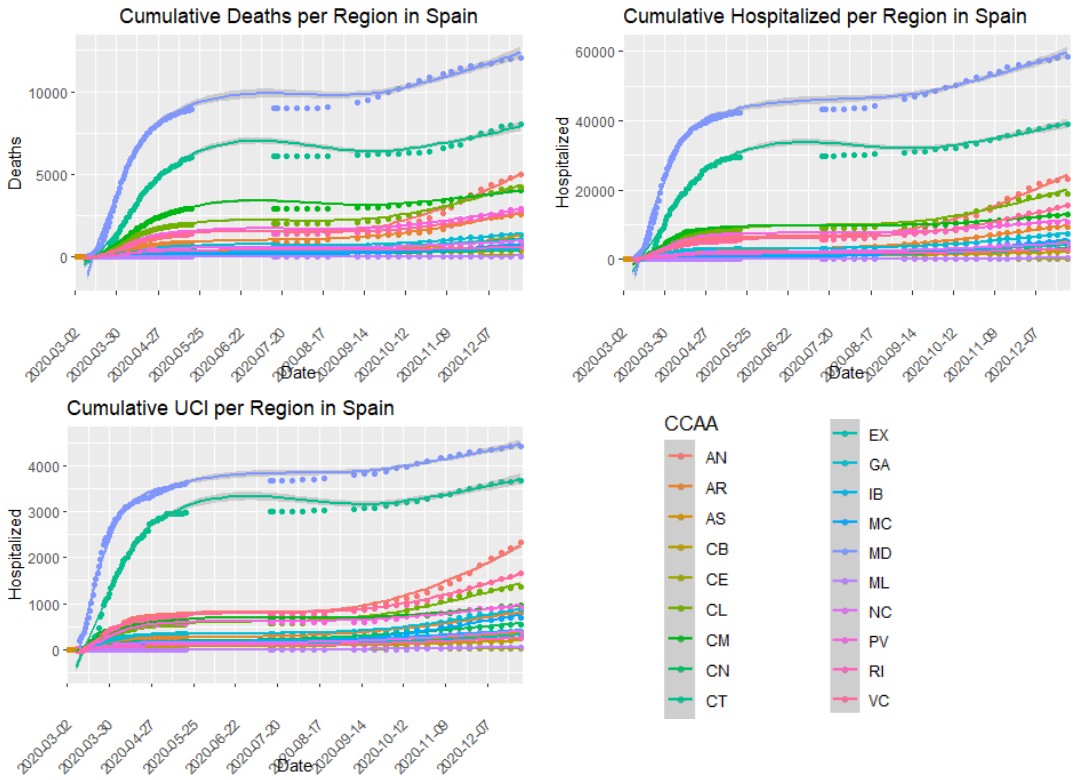
For example, the top two plots above represent the total number of male and female COVID-19 cases in Spain for all the age groups. The bottom two plots indicate the true number (percentage) of male and female COVID-19 cases after considering the population for that specific group in Spain. Future trend visualizations will be provided in subsequent reports.

The next two graphs display the percentage of COVID-19 confirmed cases for all age group distributions of the Spanish population. Some observations can be obtained from these preliminary visualizations:

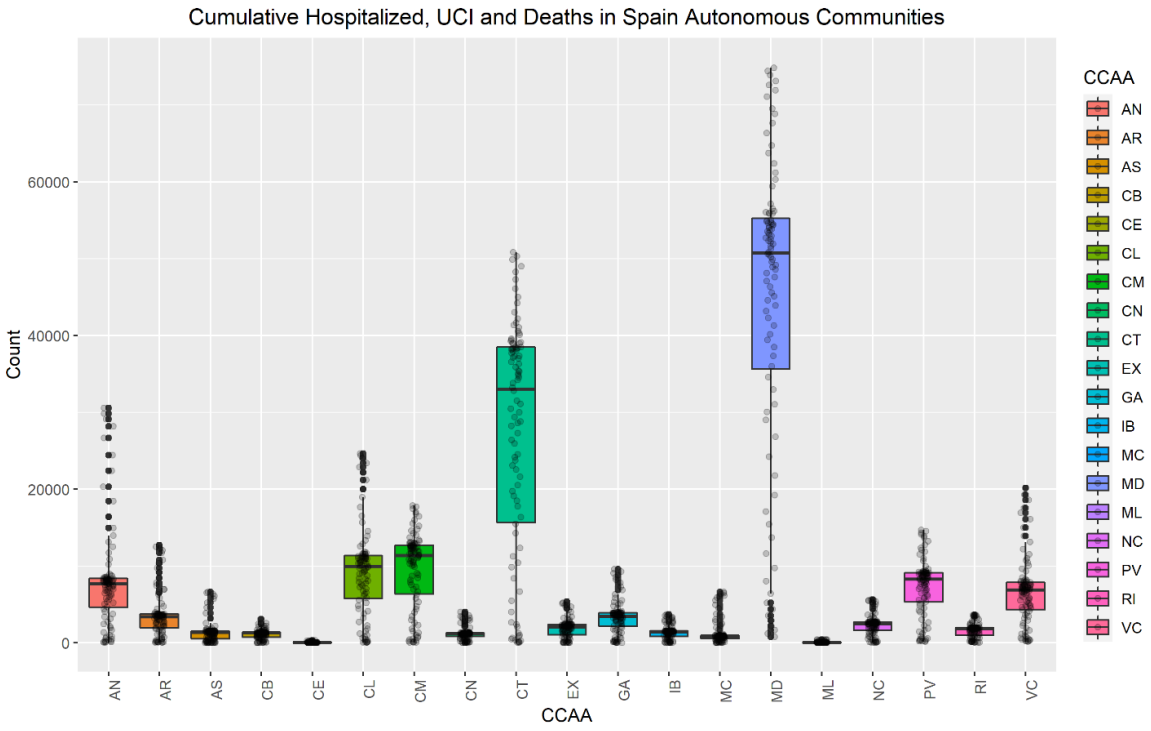
a. Higher percentage of female with COVID-19 cases for certain groups (>=80, 50-59, 40-49, 30-39 and 15-29) during early months of COVID-19 pandemic.
b. Group 70-79 shows also a higher percentage of female with COVID-19 cases but for after July whereas male show a higher rate before July.
c. Senior people (groups >=80, 70-79, 60-69, 50-59) show a higher percentage (~0.4%) during early months of COVID-19 but other groups (40-49, 30-39, 15-29, and 5-14) have a higher percentage with COVID-19 cases with above 1.5% for each respective group.



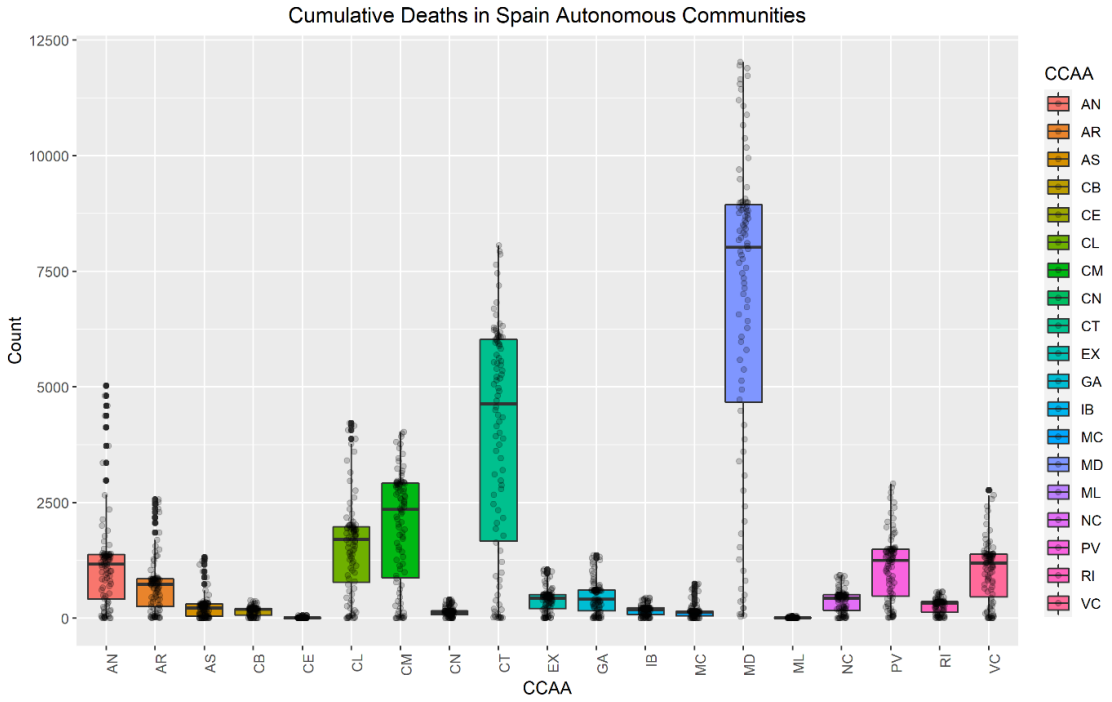
These are preliminary trends of cumulative deaths for all CCAAs in Spain:



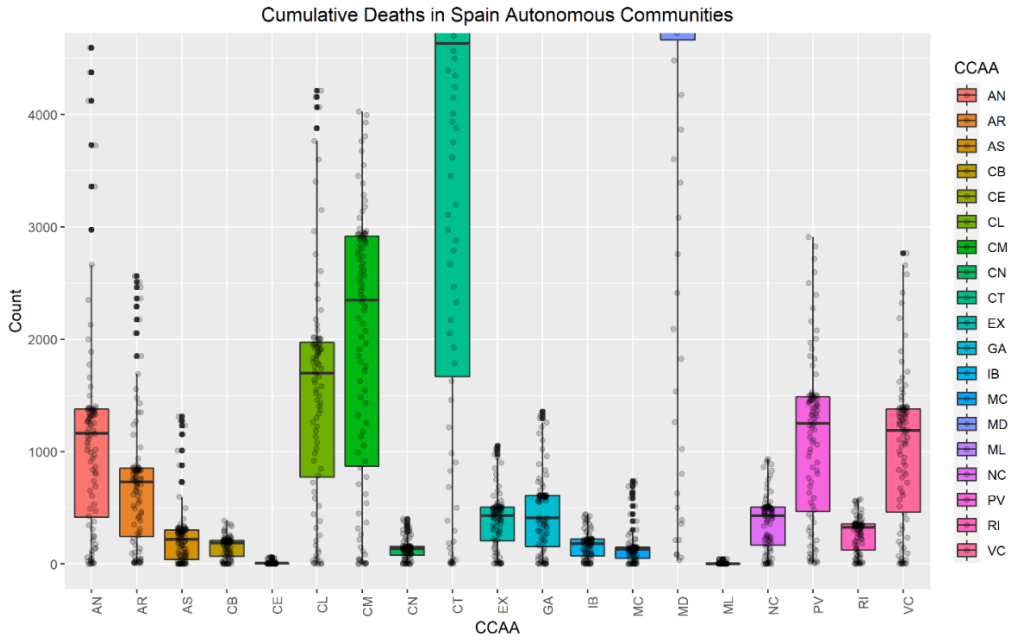
Later, we will obtain more refined trends for some prominent CCAAs, such as Madrid (MD) and Catalonia (CT).



Keeping only the number of deaths, the above plot becomes:

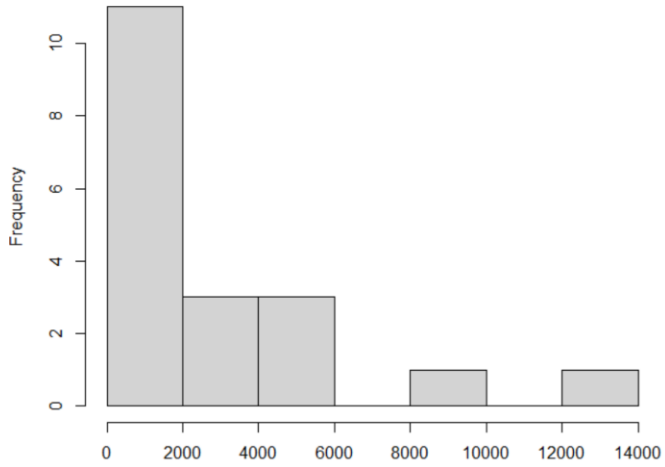


The total number of deaths is about 48872. Madrid (MD) and Catalonia (CT) account for about 41% of all deaths in Spain together. Note that this data is approximate because of the few data observations not included in this preliminary assessment. From the NIH website, Spain reached a total number of deaths of 50689, which means the algorithm processed about 96.4% of all data. Future work will include the remainder data. The next graph is a zoom of the previous one where we can see those CCAAs with lower count of deaths:



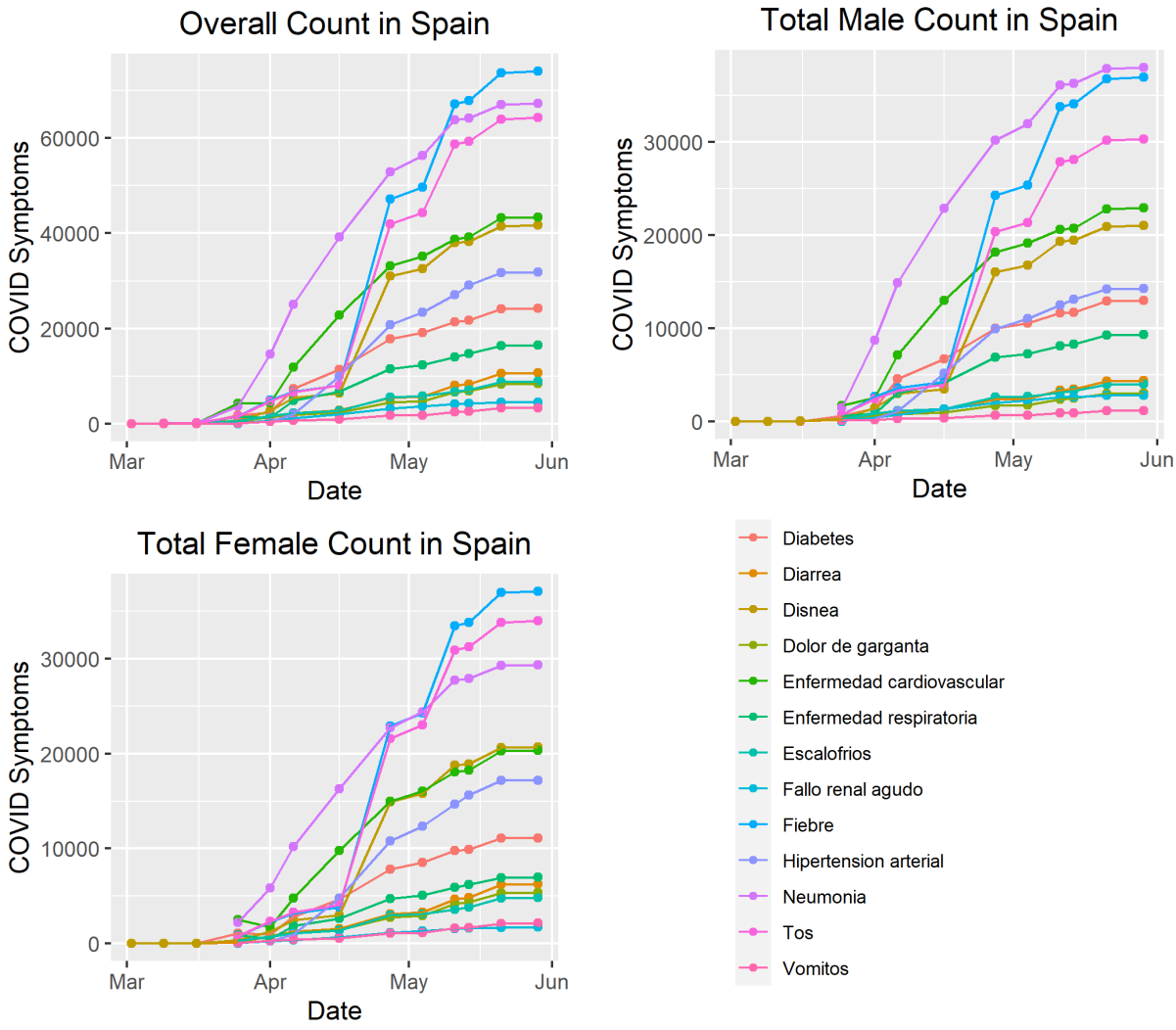
Note some key observations:

a. There are 11 CCAAs with total number of deaths up to 2000.
b. There are 3 CCAAs with total number of deaths between 4000 and 6000. Andalucía (AN), Castilla y León (CL) and Castilla la Mancha (MC). These are the largest autonomous communities in Spain.
c. There are 3 CCAAs with total number of deaths between 2000 and 4000. These CCAAs are Aragon (AR), Pais Vasco (PV) and Valencian community (VC).
d. Catalonia and Madrid are in two different brackets on their own, 8000-10000 and 12000-14000, respectively.

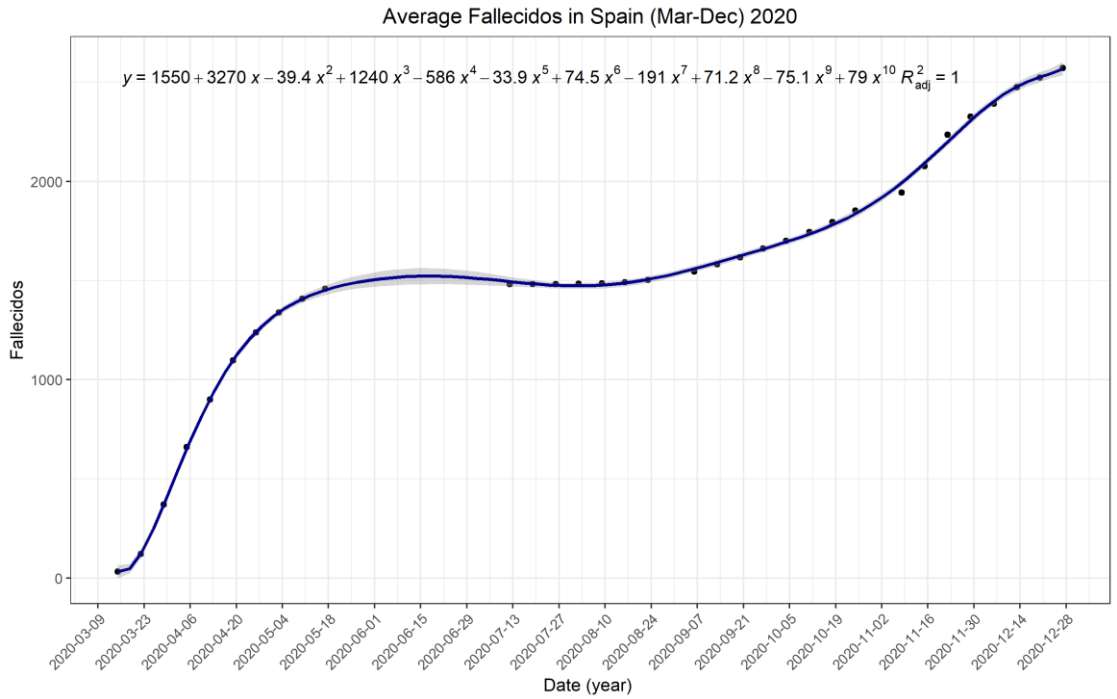


The next three graphs are quite insightful but only provided during early months of the pandemic. The RENAVE stopped providing this data after June 2020 and we will continue investigating these possible trends in the future.

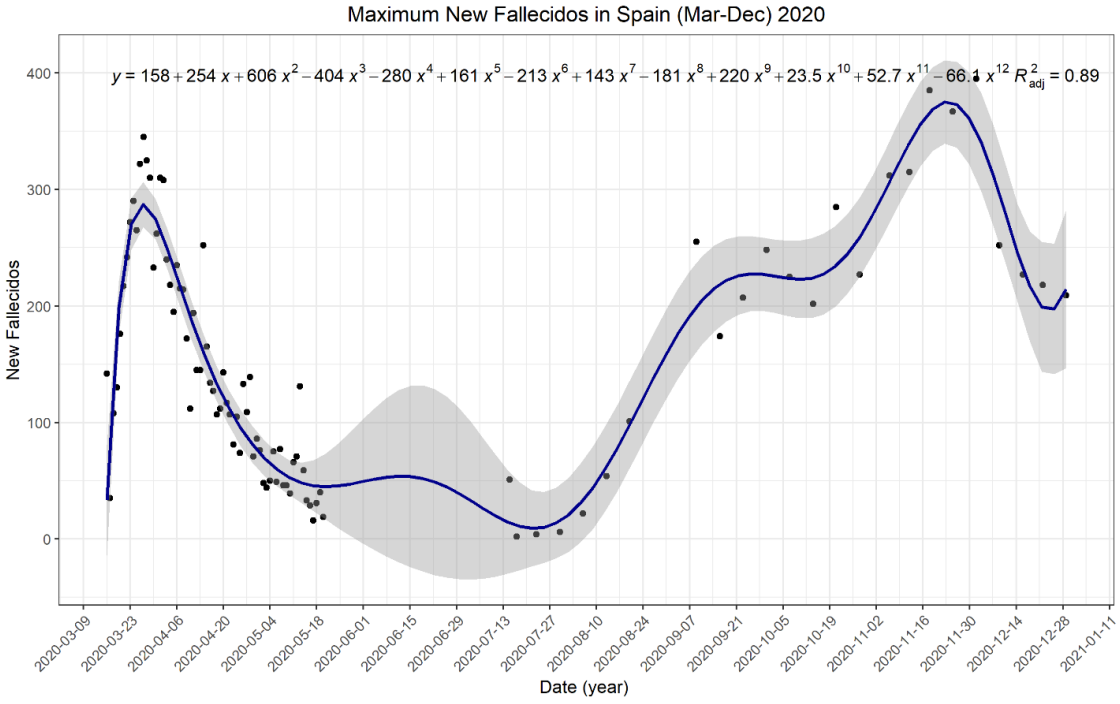
Diabetes	Diabetes
Diarrea	Diarrhea
Disnea	Dyspnea or shortness of breath
Dolor de garganta	Sore throat
Enfermedad cardiovascular	Cardiovascular disease
Enfermedad respiratoria	Respiratory disease
Escalofrios	Shaking chills
Fallo renal agudo	Acute kidney failure
Fiebre	Fever
Hipertension arterial	Arterial hypertension
Neumonia	Pneumonia
Tos	Cough
Vomitos	Vomits



It is possible to obtain a preliminary model for the average of deaths (fallecidos) in Spain. The plot below shows a trend that Spain is likely to follow in subsequent months. The most common symptoms related to coronavirus in Spain (until June) have been associated with fever, cough and pneumonia for both males and females. The next two most important symptoms are cardiovascular disease and sore throat. The next two most relevant symptoms are arterial hypertension and diabetes.

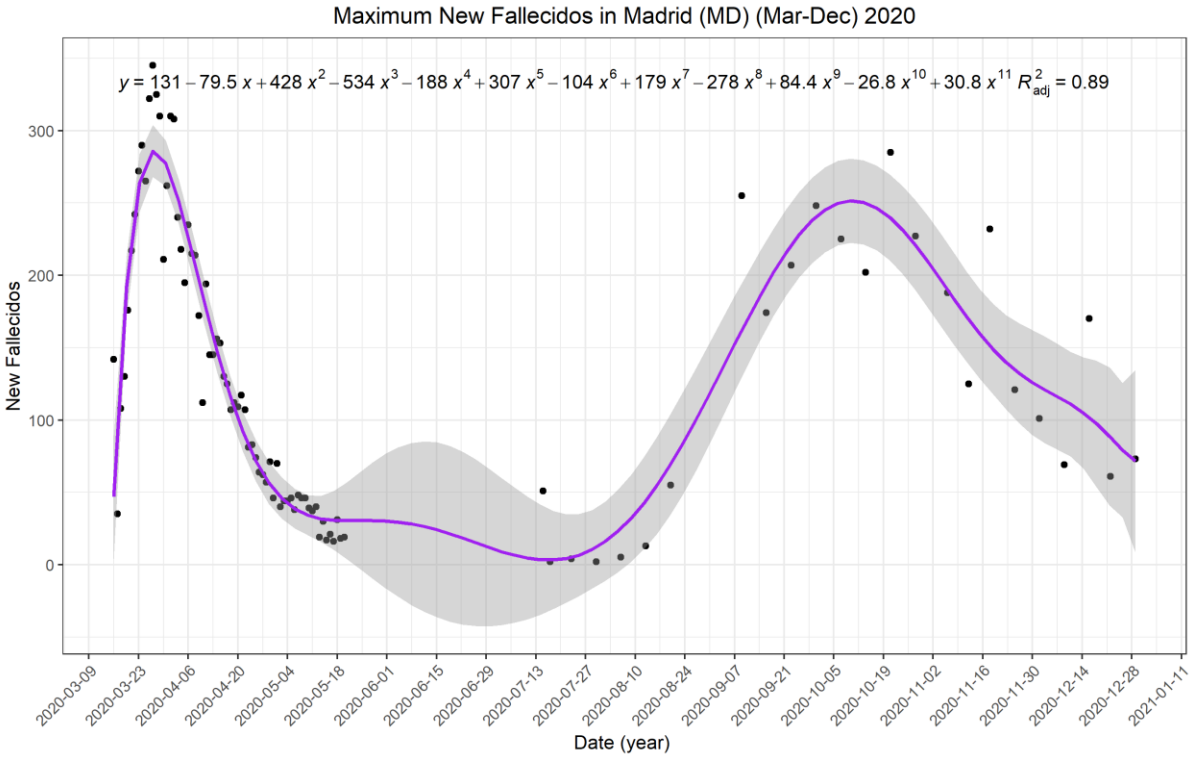


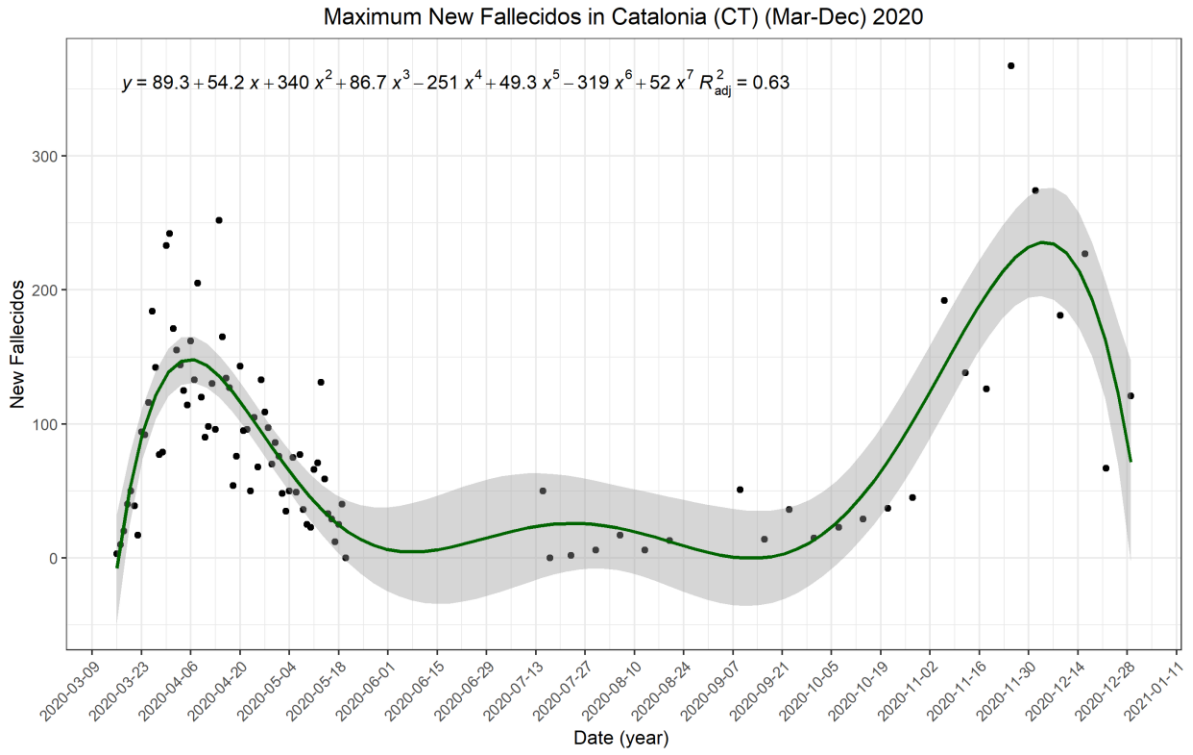
Similarly, we can obtain a profile trend of the evolution of maximum new deaths in Spain:



Note that adding all the discrete values adds up to the total number of deaths or new deaths in Spain.

As an example, we looked at the data for Madrid and Catalonia communities for which we were able to extract some preliminary models for the number of deaths.





Note that the model for Catalonia yields a rather low correlation coefficient. Although a higher correlation coefficient can be obtained (~0.69), the model fit curve goes below 0 during some time interval, which in unrealistic. Thus, this preliminary model is kept.

Similarly, we can obtain a model that explains the number of deaths or new deaths per million for each CCAA. These models vary with the number of polynomial coefficients needed in order to have a meaningful representation trend-like of number of deaths for that specific CCAA.

In this machine learning application, we selected 10% of the CoronavirusSPAIN.csv data to be the validation set. The training set is named "covid_train_set" and the test set is named "covid_test_set". Below is a summary for each of the sets:

```
> summary(covid_train_set)
     CCAA               FECHA                CASOS              PCR+           TestAc+
 Length:1723        Min.   :2020-02-29   Min.   :     0.0   Min.   :     0   Min.   :   0.0
 Class :character   1st Qu.:2020-04-02   1st Qu.:   749.5   1st Qu.:  1144   1st Qu.: 154.2
 Mode  :character   Median :2020-04-26   Median :  2243.0   Median :  3390   Median : 920.5
                    Mean   :2020-05-28   Mean   :  8065.9   Mean   : 10705   Mean   :1550.2
                    3rd Qu.:2020-07-15   3rd Qu.:  7924.5   3rd Qu.: 10606   3rd Qu.:2074.2
                    Max.   :2020-12-29   Max.   :323541.0   Max.   :255940   Max.   :8634.0
                                                            NA's   :19       NA's   :1075
 Hospitalizados         UCI           Fallecidos       New_Hosp         New_UCI        New_Fallecidos
 Min.   :    0.0   Min.   :   0.0   Min.   :    0   Min.   :   0.0   Min.   :  0.00   Min.   :  0.00
 1st Qu.:  566.5   1st Qu.:  75.0   1st Qu.:   66   1st Qu.:   6.0   1st Qu.:  0.00   1st Qu.:  1.00
 Median : 1737.0   Median : 170.0   Median :  322   Median :  31.0   Median :  3.00   Median :  7.00
 Mean   : 5612.5   Mean   : 527.7   Mean   : 1135   Mean   : 127.2   Mean   : 11.33   Mean   : 27.63
 3rd Qu.: 5948.0   3rd Qu.: 565.0   3rd Qu.: 1258   3rd Qu.: 120.0   3rd Qu.: 11.00   3rd Qu.: 28.00
 Max.   :58418.0   Max.   :4430.0   Max.   :12026   Max.   :2213.0   Max.   :242.00   Max.   :395.00
```

```
> summary(validation)
     CCAA               FECHA                CASOS             PCR+          TestAc+
 Length:117         Min.   :2020-03-05   Min.   :   5   Min.   :   4   Min.   :   0.0
 Class :character   1st Qu.:2020-03-27   1st Qu.: 256   1st Qu.: 523   1st Qu.:  51.0
 Mode  :character   Median :2020-04-17   Median : 1052   Median : 1501   Median : 440.0
                    Mean   :2020-05-09   Mean   : 2043   Mean   : 3293   Mean   : 916.3
                    3rd Qu.:2020-05-16   3rd Qu.: 2186   3rd Qu.: 2772   3rd Qu.:1001.0
                    Max.   :2020-12-09   Max.   :18929   Max.   :70841   Max.   :6285.0
                                                                         NA's   :72
 Hospitalizados       UCI          Fallecidos       New_Hosp         New_UCI         New_Fallecidos
 Min.   :    2   Min.   :   1.0   Min.   :   0.0   Min.   :   0.00   Min.   :  0.000   Min.   :  0.00
 1st Qu.:  205   1st Qu.:  25.0   1st Qu.:  14.0   1st Qu.:   2.00   1st Qu.:  0.000   1st Qu.:  0.00
 Median :  822   Median : 103.0   Median : 115.0   Median :  11.00   Median :  1.000   Median :  2.00
 Mean   : 1569   Mean   : 160.8   Mean   : 312.9   Mean   :  66.28   Mean   :  6.932   Mean   : 11.23
 3rd Qu.: 1775   3rd Qu.: 160.0   3rd Qu.: 399.0   3rd Qu.:  48.00   3rd Qu.:  6.000   3rd Qu.:  7.00
 Max.   :14918   Max.   :1532.0   Max.   :2962.0   Max.   :1528.00   Max.   :131.000   Max.   :217.00
```

For the Spain study, the mean of total deaths was found to be $\hat{\mu}$=1135.044 with a RMSE of 968.6117 using the average method. The code saves the RMSE table as RMSE.csv file, which is the table below:

method	RMSE
Just the average	**968.6117**
CCAA Effect Model	839.4386
CCAA + Hospitalizados Effects Model	302.2603
CCAA + Hospitalizados + UCI Effects Model	246.0710
CCAA + Hospitalizados + UCI + FECHA Effects Model	247.6527
Regularized CCAA + Hospitalizados + UCI + FECHA Effects Model	241.5004

Some take away notes:

- We see a decrease of about **13.3%** in the RMSE of "CCAA Effect Model" with respect the "Just the average" method.
- The RMSE was decreased by about **64%** when using the "CCAA + Hospitalized Effects Model" with respect the "CCAA Effect Model".
- The RMSE was slightly lowered by **18.6%** when using the "CCAA + Hospitalized + UCI Effects Model" with respect to the "CCAA + Hospitalized Effects Model".
- The RMSE obtained using the "CCAA + Hospitalized + UCI + FECHA Effects Model" and the "CCAA + Hospitalized + UCI Effects Model" is almost the same.
- Finally, regularization decreased the RMSE by about **2.5%** with respect to the " CCAA + Hospitalized + UCI + FECHA Effects Model".
- Adding UCI and FECHA effect in models 3 and 4 add very small gains in terms of model performance and model 5 (regularization) happens to perform better. The lowest RMSE was about 21.3% as large as the mean.

For the Europe study (data provided every day), the mean of total deaths per million was found to be $\hat{\mu}$=191.9184 with a RMSE of 178.7406 using the average method. The RMSEs obtained for the Europe study are:

method	RMSE
Just the average	**178.74059**
New Cases Per Million Effect Model	111.83848
New Cases Per Million + Location Effects Model	10934776
New Cases Per Million + Location + Date Effects Model	105.34562
Location Effect Model	106.79341
Location + Date Effects Model	81.32859
Location + Date + New cases Per Million Effects Model	66.54128
Date Effect Model	145.14379
Date + Latitude Effects Model	78.20350
Date + Longitude Effects Model	78.20350
Date + New Cases Per Million Effects Model	90.56084
New Cases Per Million + Date Effects Model	88.62267
Positive Rate Effect Model	154.76913
Positive Rate + New Cases per Million Effects Model	125.96897
Regularized Date + New Cases Per Million Effect Model	89.04313
Regularized Positive Rate Effect Model	151.32458
Regularized Positive Rate + New Cases Per Million Effect Model	118.56549
Regularized New Cases Per Million + Date + Positive Rate Effect Model	94.57055
Regularized Date + New Cases Per Million + Positive Rate Effect Model	88.05126

Some take away notes from the Europe study case:

- Location, longitude and latitude generate the lowest RMSE among all models. However, latitude and longitude for all European countries was centered in each country geographically. Thus, these results do not consider variations of observations beyond these locations, which could alter the RMSES significantly. We saw that for the Spain study, some regions are more dominant than other. Even if we keep the lowest RMSE using this preliminary model, it is still half of the mean.
- The positive rate variable generates the largest RMSEs values than other models tested.
- Regularization slightly reduces the RMSE for each respective model with same variables.
- Considering certain groups of countries or neighbor countries in Europe may be an alternative to better understand and interpret these models. It is expected that doing so, the RMSE would be decreased significantly. This is currently being studied but it is not provided in this report.

IV. CONCLUSIONS and FUTURE WORK

The goal of this preliminary work was to understand the coronavirus spread across Europe and about Spain using data science and machine learning tools. Some of these results are unique and have not been published elsewhere. For simplicity, not all visualizations that the algorithm generates were included in this preliminary report.

Future work can involve analyzing some of these CCAAs together including travel between CCAAs for the Spain study. For example, from the above histogram we can see that three groups can be studied separately to improve the predictive model(s). For example, one model could be just Madrid and Catalonia, the second model can include 6 CCAAs with deaths between 2000-6000, and the last model may involve 11 CCAAs with less than 2000 deaths. These results will be provided in a subsequent report. In our study, the most accurate model predicting the number of deaths in Spain was about 1135 ± 242 per interval of data obtained (per week). From early March until end of December of 2020 there were 43 weeks. Thus, 1135*43 = 48805 deaths, which is close to the approximated value, 48872 deaths processed by our algorithm and provided by the Spanish RENAVE. However, this number as mentioned earlier, is expected to be around 5% higher.

For the Europe case, it seems a bit more complex given that large number of variables, uncertainty of observations due to geographical locations of countries, and lack of knowledge of travel between large main cities within countries. Several models yield RMSE larger than the mean. To improve these results, we would need to have a better understanding on the effect of some of the above variables for a few countries at a time geographically located, and travel considerations between and within these countries. Because the coronavirus is an evolving virus, it constantly affects the countries in different manner yielding different distributions for the number of deaths per million for each country, thus the RMSE can be large. In our study, we obtained an average of 226 deaths per million for the European

countries. For 43 weeks or about 300 days, it is estimated approximately 300*(192-(RMSE=151))= 12300 deaths in average for each of these countries. We considered the RMSE of the regularization for the positive rate. Since there are 46 countries, then 12300 *46 = 565800 total deaths, which is an overestimation (~3.8% higher) from the actual approximated number of deaths in Europe or 545150 deaths in 2020. Assuming a model (date) with lower RMSE=145 , then the estimated total of deaths in Europe for 2020 would be 648600, which is an overestimation of the actual value by ~19.0%.

Other machine learning and data science tools will be implemented to refine such models and have an improved knowledge and understanding of the coronavirus that can leverage the efforts of epidemiology entities. Publication of some results are currently being considered. Collaboration on this topic are welcome and encouraged.

Future versions of this R-algorithm will be further refined and optimized.

V. REFERENCES

[1] Introduction to Data Science: Data Analysis and Prediction Algorithms with R. Front Cover. Rafael A. Irizarry. CRC Press, Nov 20, 2019.

[2] Open-Access Data and Computational Resources to Address COVID-19**,** https://datascience.nih.gov/covid-19-open-access-resources, (accessed December 31 2020).

[3] Centro Nacional de Epidemiología, https://covid19.isciii.es/ (data retrieved on December 29 2020).

[4] Population Pyramids of the World from 1950 to 2100, https://www.populationpyramid.net/spain/2019/.

[5] Countries Latitudes and Longitudes, https://developers.google.com/public-data/docs/canonical/countries_csv

VI. APPENDIX

The R-algorithm used to generate this report and data can be found in the GitHub repository under https://github.com/PedroLlanos