26-6-2022

# COVID-19 Global Forecasting

## Midterm Project

ALEJANDRO CARMELO PARRA GARCIA
PREDICTIVE MODELS

# Table of Contents

## Problem

The Covid-19 virus and the consequent pandemic broke into the world quickly and with very serious consequences, causing a large number of deaths and illnesses, blocking and paralyzing economies, countries and entire continents.

For this project I'm trying to predict or forecast future covid-19 daily cases of various locations.

## Significance

If I'm able to build models that reliably predict future covid-19 cases I will be able to anticipate the pandemic, being able to plan resources more efficiently. By knowing when we will get a spike in the number of cases, we can dedicate more resources to hospitals in order to treat all patients.

## Data Description

The Time series is divided between countries, having a total of 187 different ones, from Afghanistan to Zimbabwe.

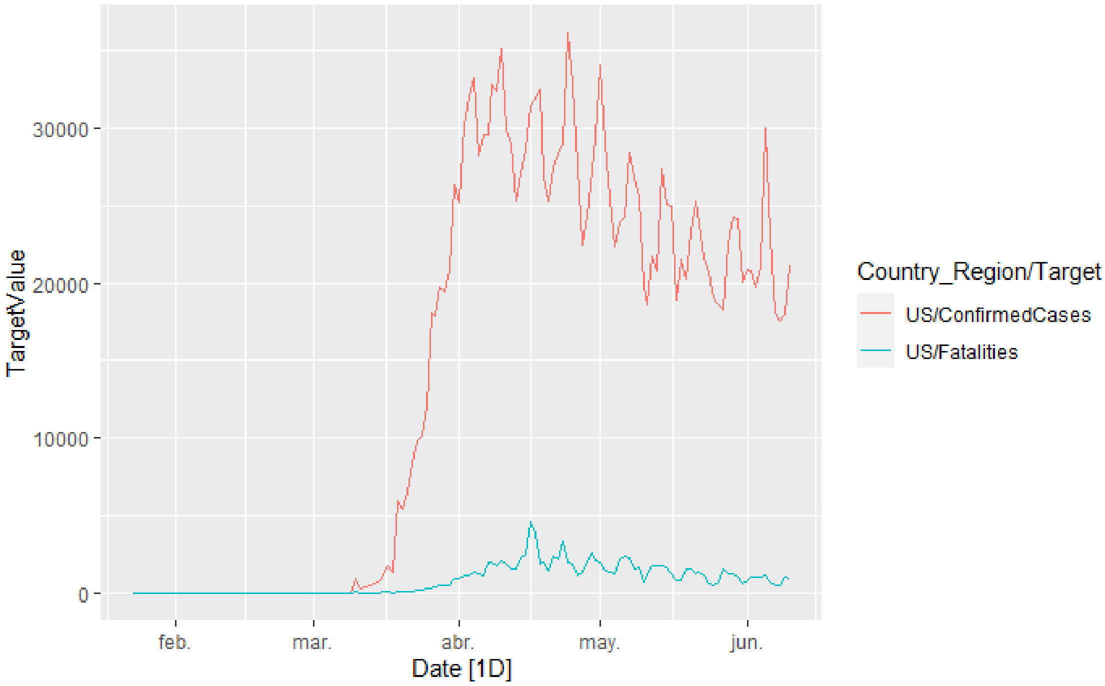
But some of these countries are divided into provinces, that being the case for 8 countries: Australia, Canada, China, Denmark, France, Netherlands, United Kingdom and US. For each of these ones we have a dataset for the hole country as well as for his provinces. The one with the highest number of provinces is the United States with 54, this is because it is including the District of Columbia, Guam, Puerto Rico and the Virgin Islands. The second country with the most provinces in China with 33, followed by Canada with 12, France and the UK with 10, Australia with 8, the Netherlands with 4 and Denmark with 2.

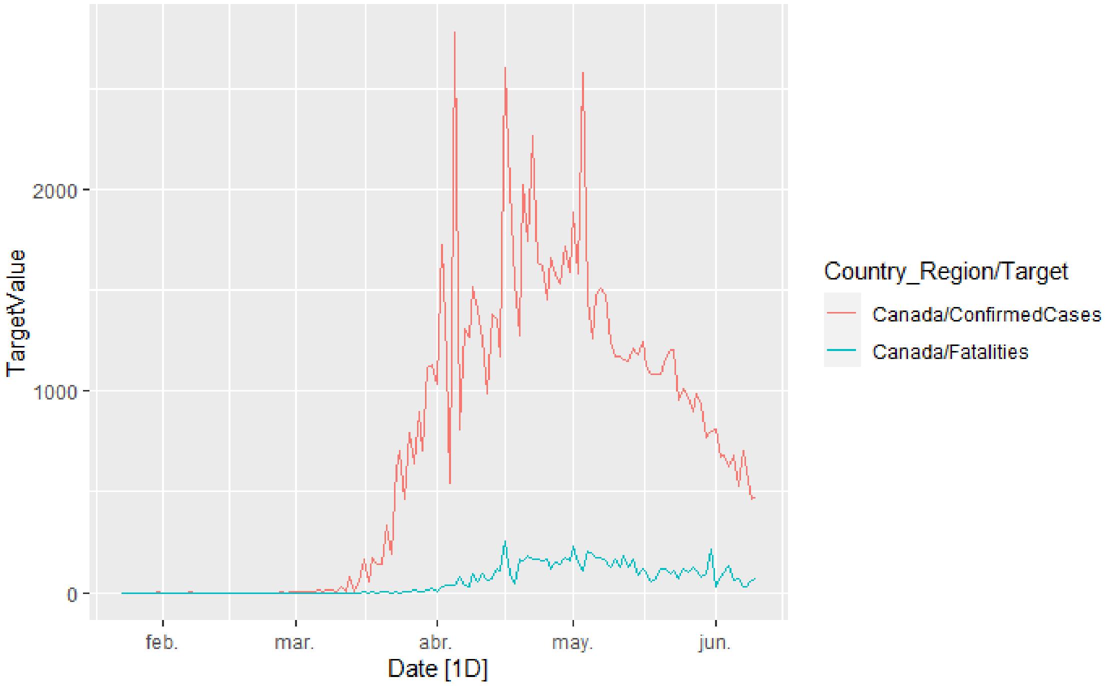
The last territorial subdivision is counties, there are 1841 different ones. All of them belongs to the US, the list of provinces or states with subdivisions has a size of 51, this is because in this case Guam, Puerto Rico and the Virgin Islands do not have counties.

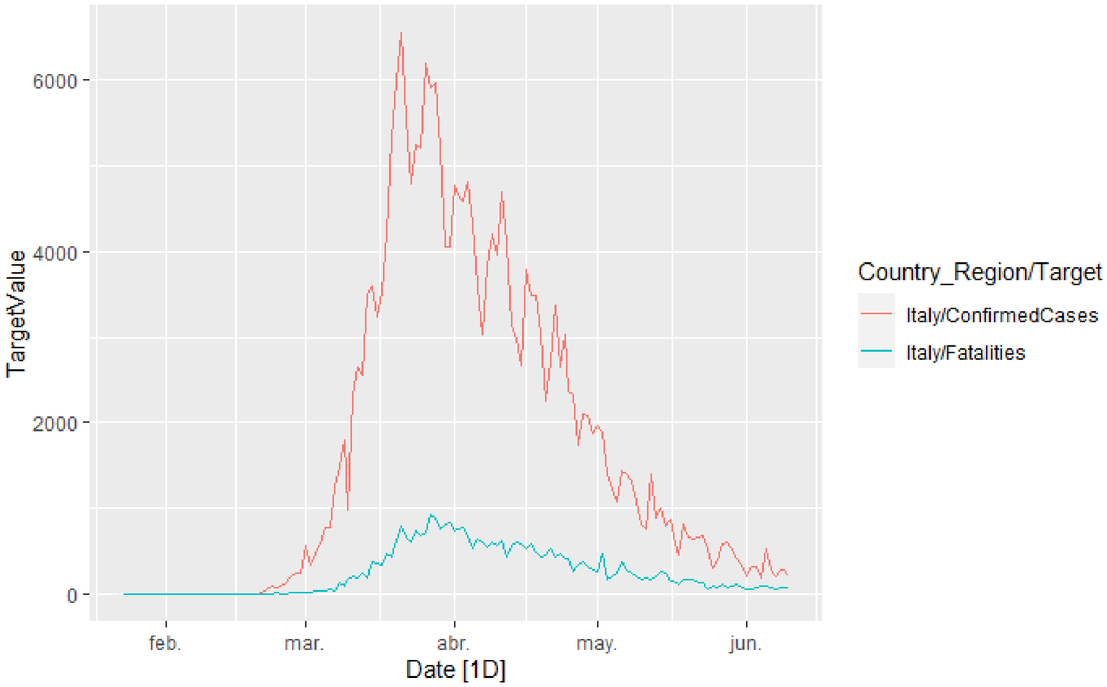
Lastly for each of these countries, provinces, counties we have two types of data, that being Confirmed Cases and Fatalities.

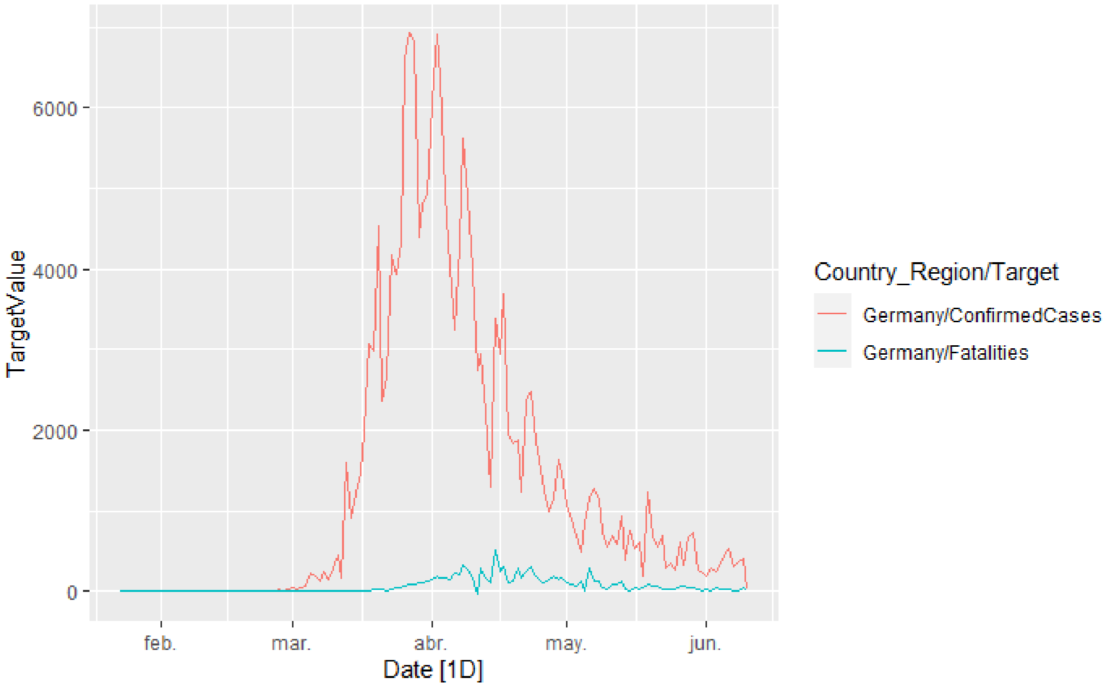
I decided to plot some countries to see how the data behaves, I choose US, Canada, Italy, Germany:

First a normal line plot to see the evolution of the data:


TargetValue
30000
20000
10000
0
feb.
mar.
abr.
may.
jun.
Date [1D]
Country_Region/Target
US/ConfirmedCases
US/Fatalities


TargetValue
2000
1000
0
feb.
mar.
abr.
may.
jun.
Date [1D]
Country_Region/Target
Canada/ConfirmedCases
Canada/Fatalities

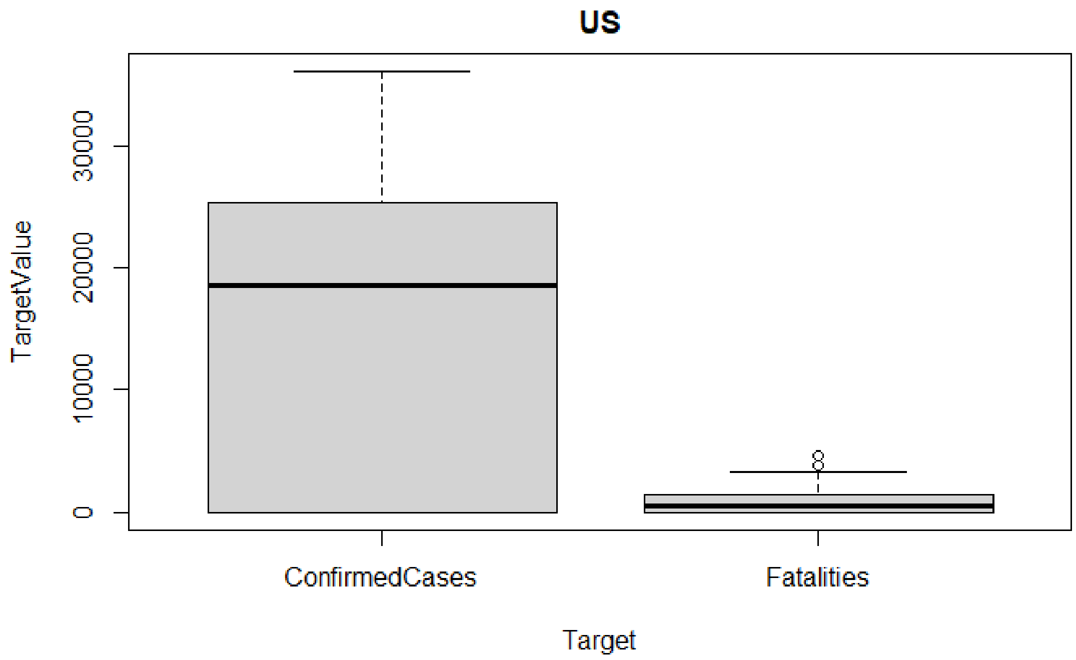


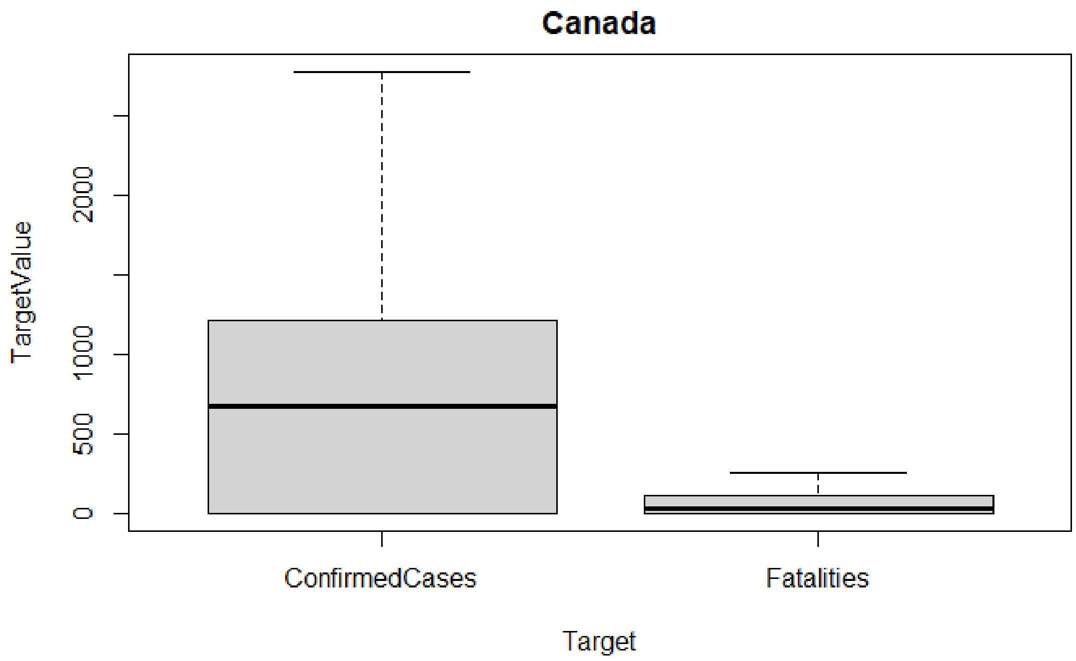


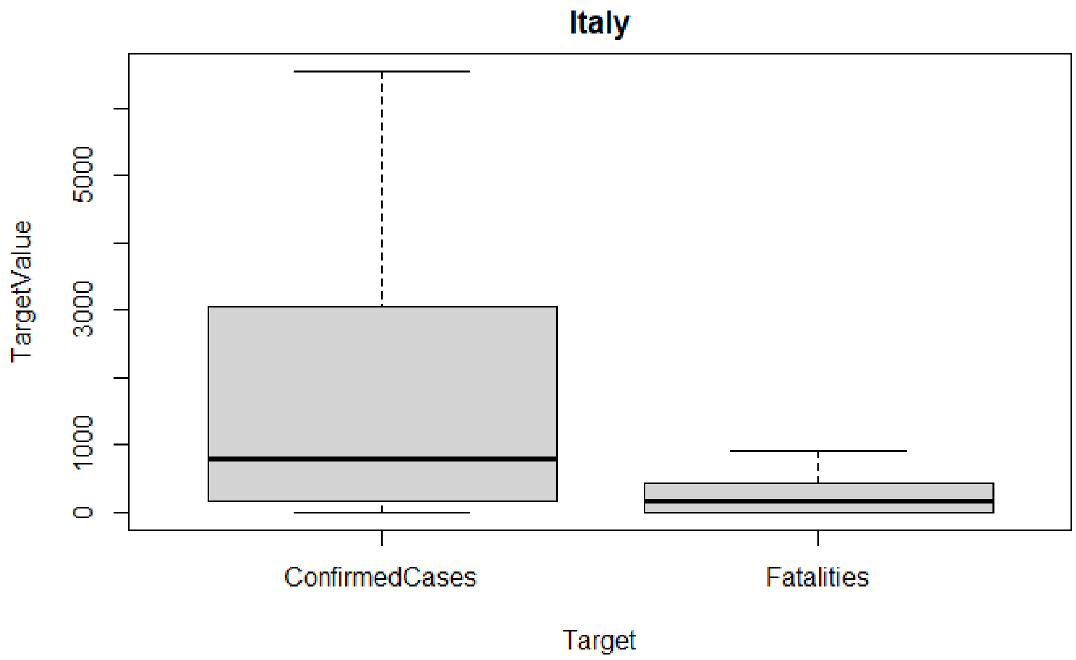
We can see that the confirmed cases datasets have higher values than the fatalities one, that's expected, since the deaths needs to be infected beforehand, but being infected doesn't means death.

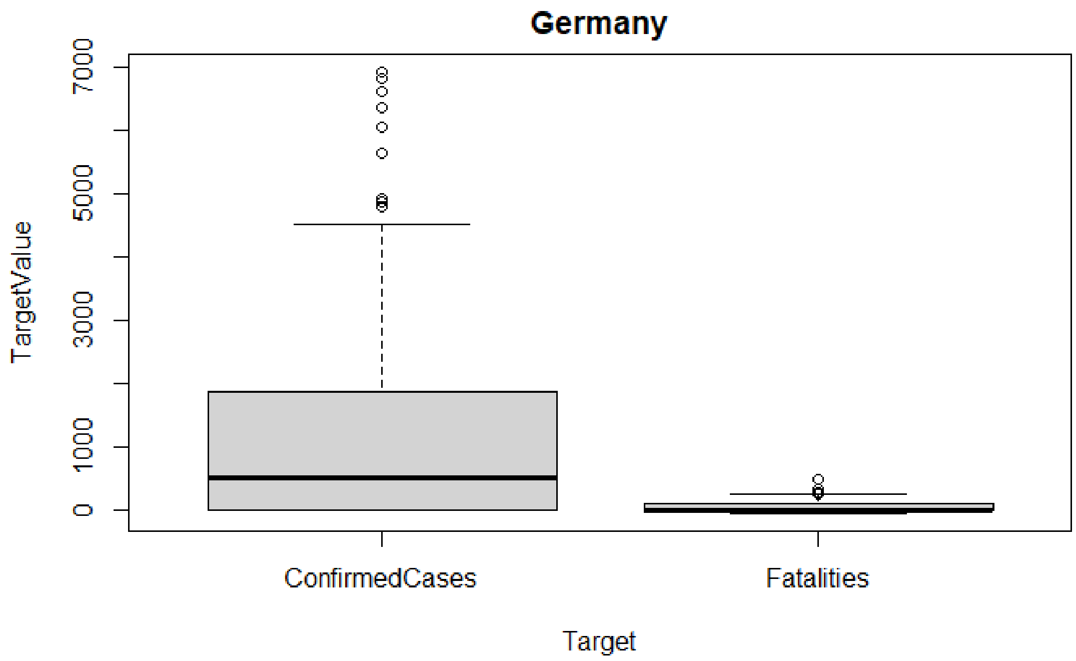
We can also see some short of waive pattern in all the datasets.

Then I plotted a normal boxplot:


US
TargetValue
0
10000
20000
30000
ConfirmedCases
Fatalities
Target

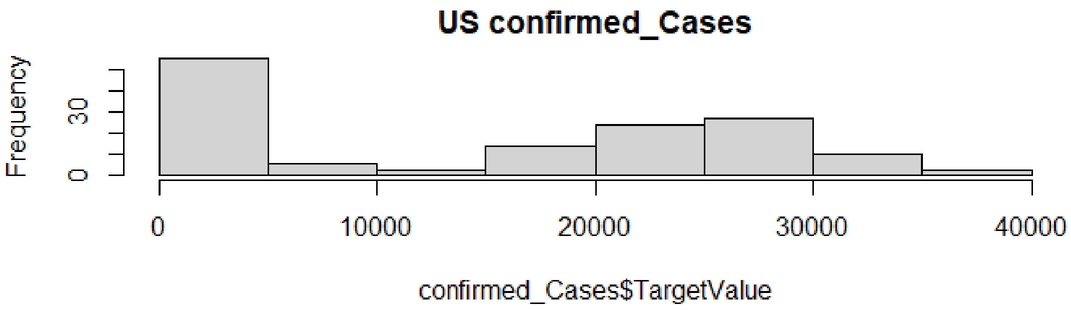

Canada
TargetValue
0
500
1000
2000
ConfirmedCases
Fatalities
Target

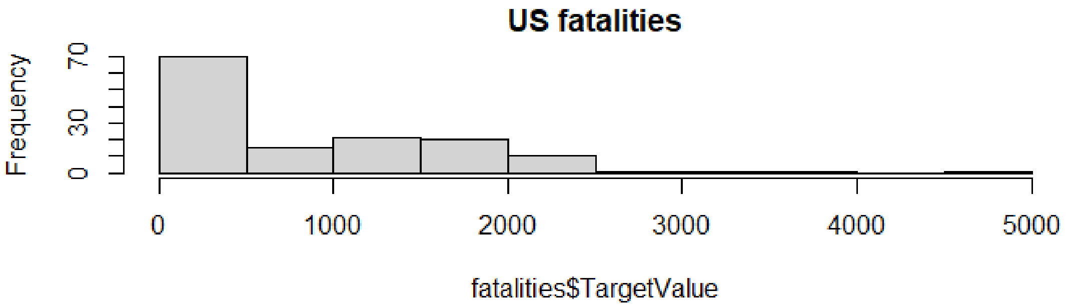


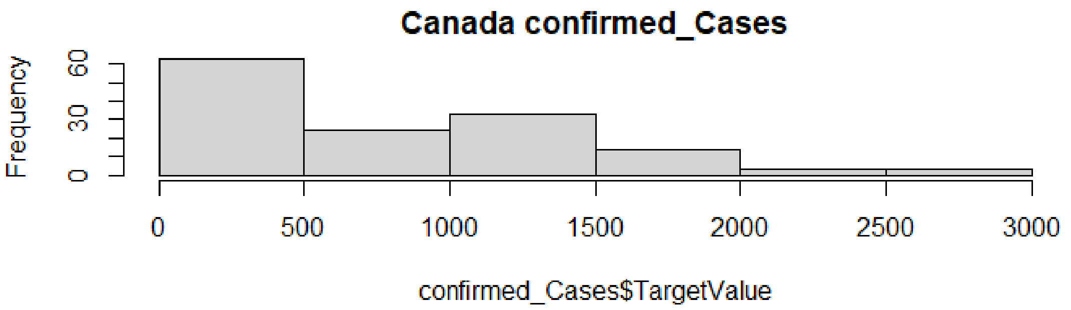


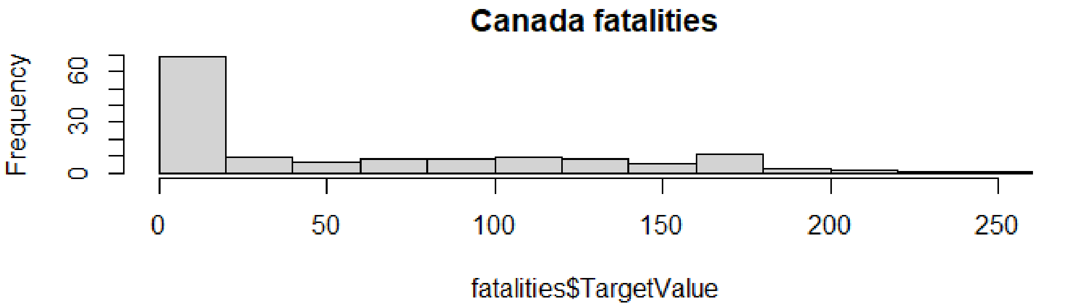
As above we can see that fatalities are closest to the 0 value than confirmed Cases. We also see that the bulk of the values for confirmed cases are close to the 0 mark, this is because the first months of the year 2020 there weren't infected people in most of the countries.

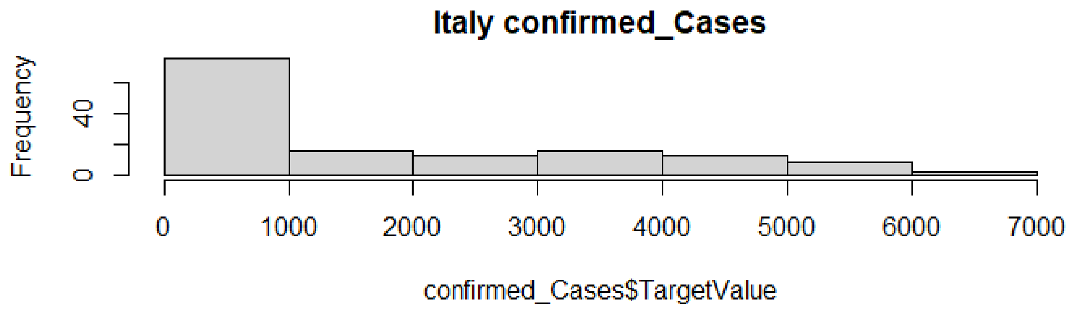
Lastly the histogram of these countries differentiating by confirmed cases and fatalities:

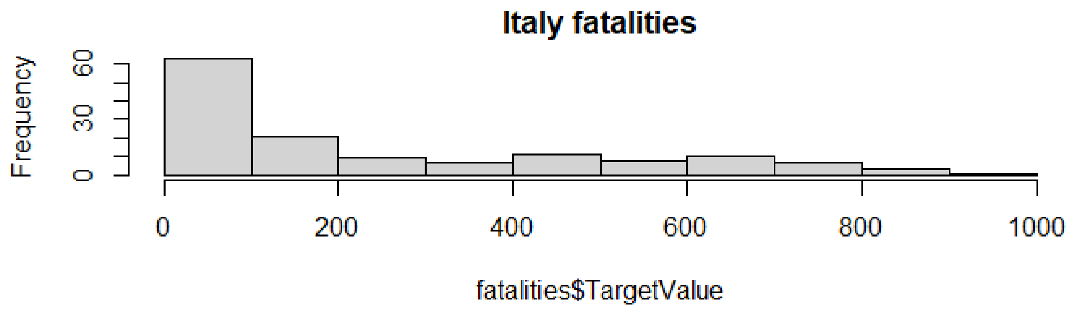

US confirmed_Cases
Frequency
0
30
0
10000
20000
30000
40000
confirmed_Cases$TargetValue

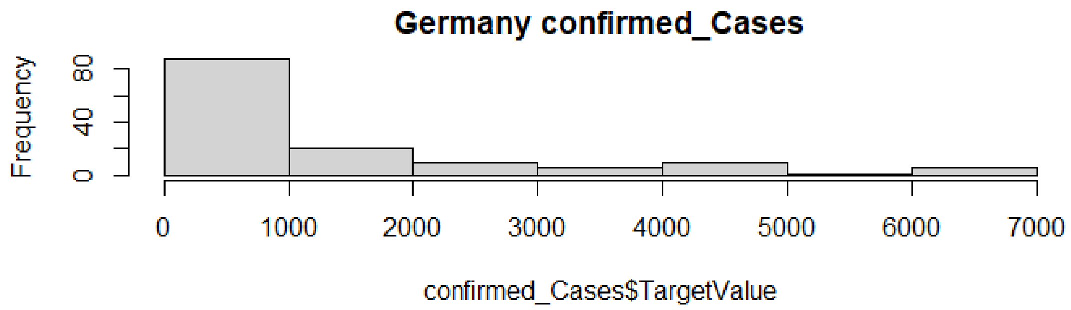

US fatalities
Frequency
0
30
70
0
1000
2000
3000
4000
5000
fatalities$TargetValue

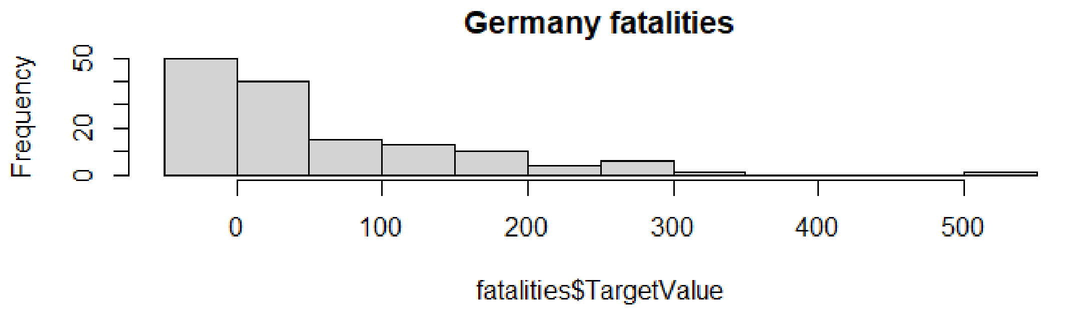

Canada confirmed_Cases
Frequency
0
30
60
0
500
1000
1500
2000
2500
3000
confirmed_Cases$TargetValue


Canada fatalities
Frequency
0
30
60
0
50
100
150
200
250
fatalities$TargetValue





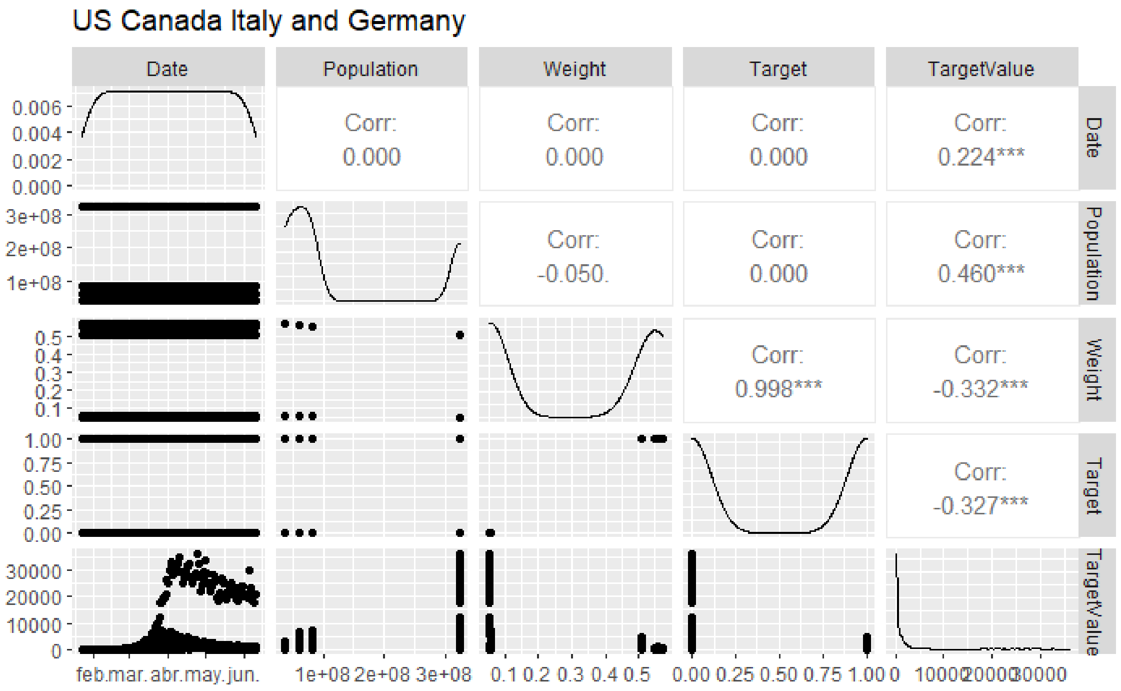


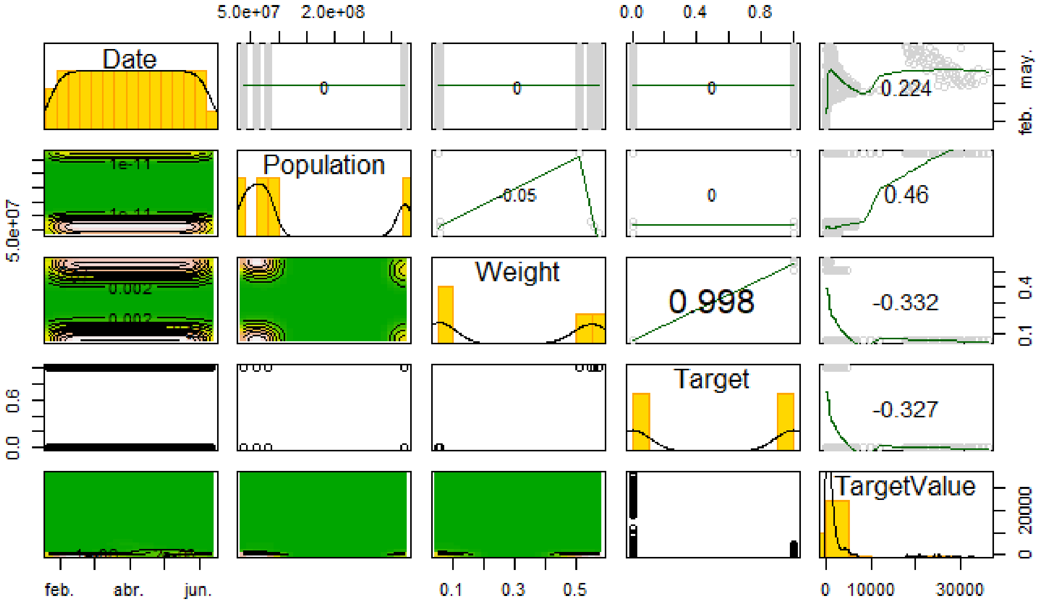


We can see a big number of the values really close to the 0 mark, this is because of the same reason as above, where the covid crisis started at different times in each country. Therefore, it will be interesting to add a dummy variable in order to represent this difference.

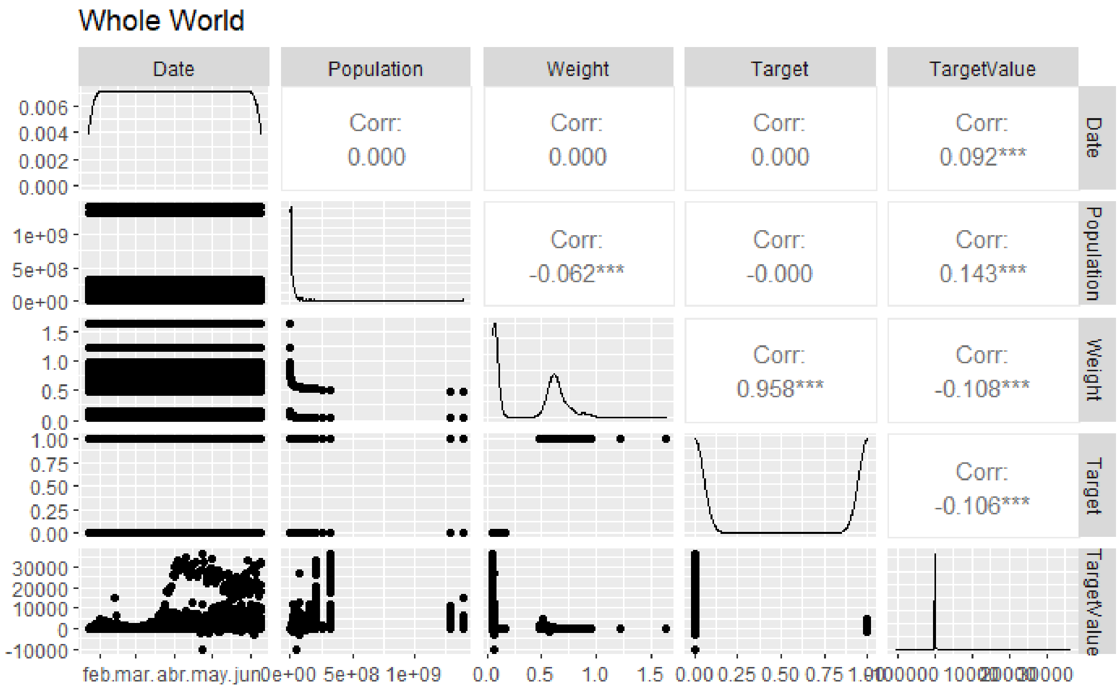
Now let's get the Scatterplot Matrix:

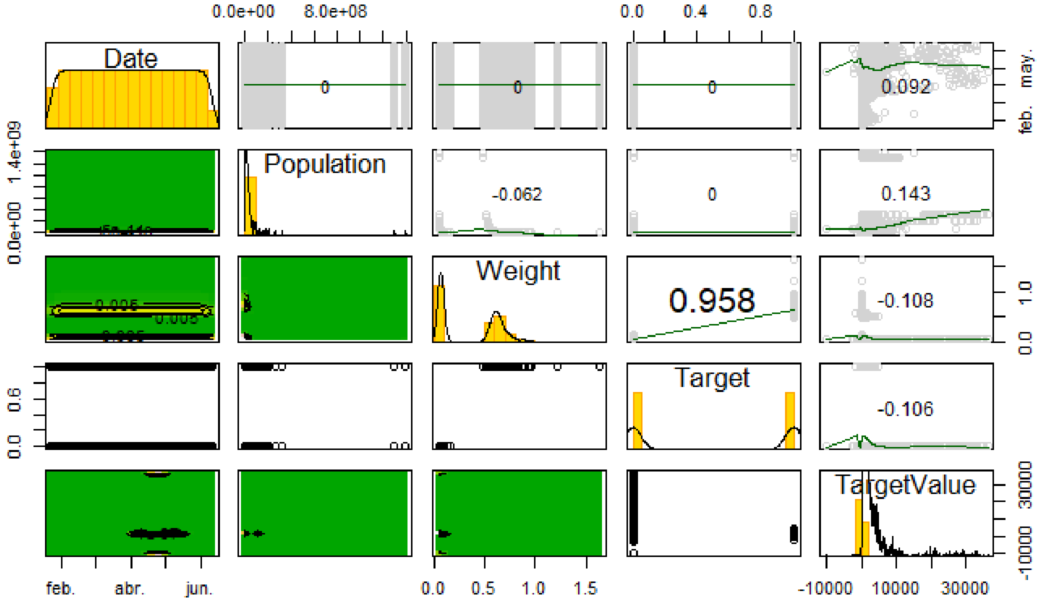
First for the 4 countries we were exploring before, US, Canada, Italy and Germany



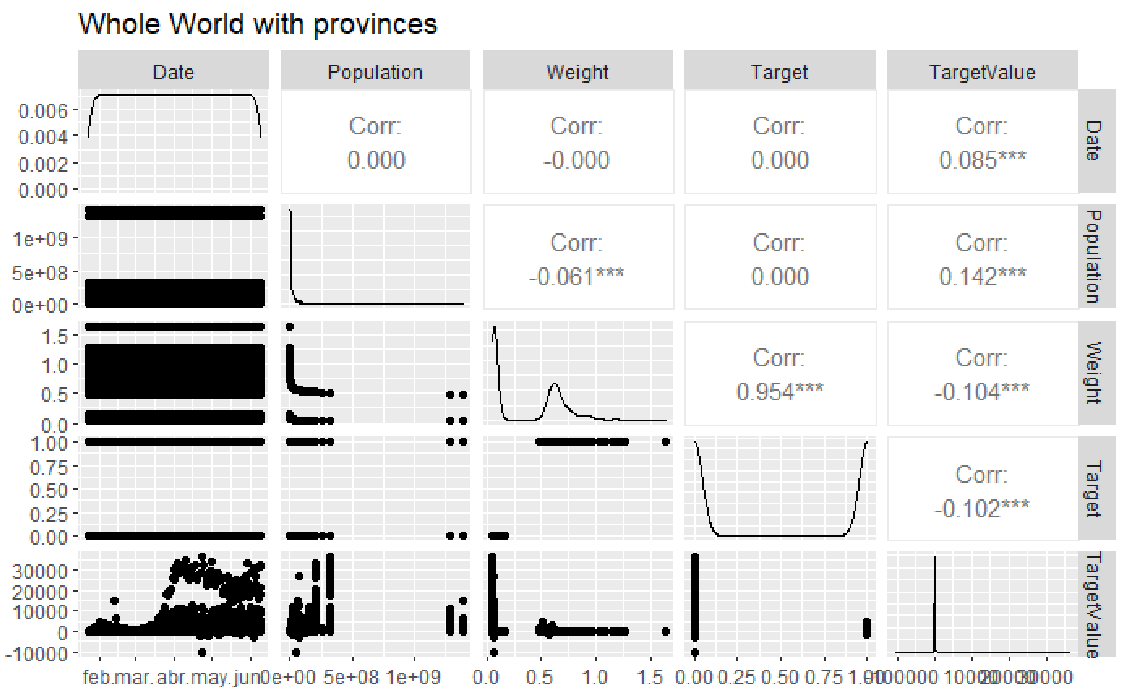


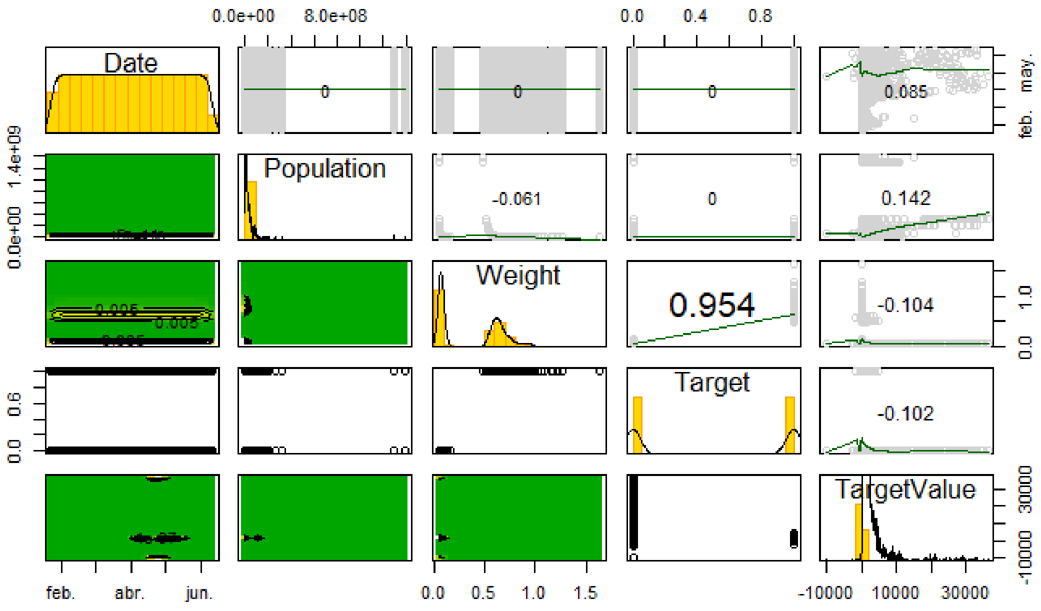
For all the countries:





Now including the provinces on top of the countries:





We can see that all of the variables are uncorrelated except for Weight and Target, this is because the weight is calculated using one formula or another depending on the target variable, that is why they are mostly correlated.

## Types of Models

The typical model will be a linear regression, although this will fail to capture the true nature of the data, as we have seen in the graphs the data does not follow a straight line.

We can use a multiple linear regression using multiple inputs, such as population, weight, or outside factors such as country GDP. This short of models would be able to predict with a higher precision.

It will be interesting to add other predictors in order to try to increase the accuracy of the model. This predictor could be the trend in order to capture the time dimension into the predictions. also using dummy variables such as one marking the start of the pandemic for each country, since the start date differs by country, meaning that for an initial period all values will be 0 in the target variable, also trying to apply seasonality, since all the dataset is contain within one year we cannot use the month as seasons, it will be wiser to use days of the week, this is important since some countries do not offer data on a daily basis.

We could also apply transformations to the predictors and or the forecast variables, such as logarithms, if we apply it to both sizes, it’s called log-log, if we apply it to only the forecast variable it is log-linear, and if it is applied only to the predictors is linear-log. In this case, if we want to apply the log functions to the forecast variable, we first will need to change the negative values to 0 and use log(x+1) in order to take into account, the zeros.

Another interesting model will be autocorrelation, we can give as predictors previous values of the dataset, since todays values are influenced by previous ones

## Literature

From reading the book and some of the discussions on the Kaggle competition we are able to get some insight into other people approaches.

For the thread about sharing datasets, we are able to find some complete dataset containing not only the original information but even more, for example from the user “Sadhaklal” (https://www.kaggle.com/code/sambitmukherjee/covid-19-data-adding-world-development-indicators/data?select=indicator_dictionary.csv), he compose a dataset including the World Bank world Development Indicators, including things like GDP in purchasing power parity (PPP), when comparing different countries from different socioeconomics factors it is better to use PPP than normal nominal GDP, since it takes into account the different prices of needs in the different economies. We can also get access to other variables like access to electricity, passengers carried, etc.

Another interesting source is the user CPMP thread about “Linear Regression Is All you Need” (https://www.kaggle.com/competitions/covid19-global-forecasting-week-5/discussion/151461), where he uses auto regression, using past values as input. As well as using the logarithm of the variables instead of the normal variables.

# Models

For the models that I have built first a normal model using only trend and season, lag weight and Pandemic Start in order to make the forecast. The 'PandemicStart' variable was created in as a dummy or Intervention variable, its value is 0 from the start until the first case is diagnosed, after that then it changes to 1 until the end. I decided to create this variable because for some countries the pandemic starts after a while, meaning that for the initial period the values of the series are 0.

The data is split into 80% for train (Before 2020-05-14) and 20% to validate (After and including 2020-05-14).

This are the models I built, for each of the models that used the lag parameter, I built multiple ones with different lag, 1, 2, 4, 6 and 8. The exceptions are the log-log models where the lags I tested are 1, 2, 4 and 6. This resulted in 37 different models.

TargetValue ~ Trend + season

TargetValue ~ Trend + season + Weight

TargetValue ~ lag(TargetValue)

TargetValue ~ Trend + lag(TargetValue)

TargetValue ~ Season + lag(TargetValue)

TargetValue ~ Trend + season + lag(TargetValue)

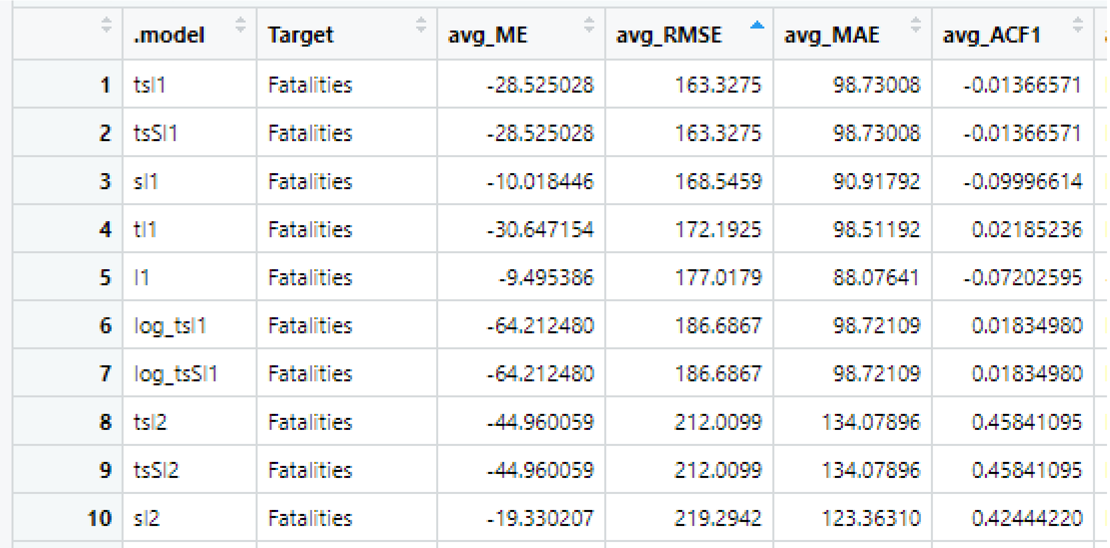
TargetValue ~ Trend + season + PandemicStart

TargetValue ~ Trend + season + PandemicStart + lag(TargetValue)

log(TargetValue) ~ Trend + season + lag(log(TargetValue))

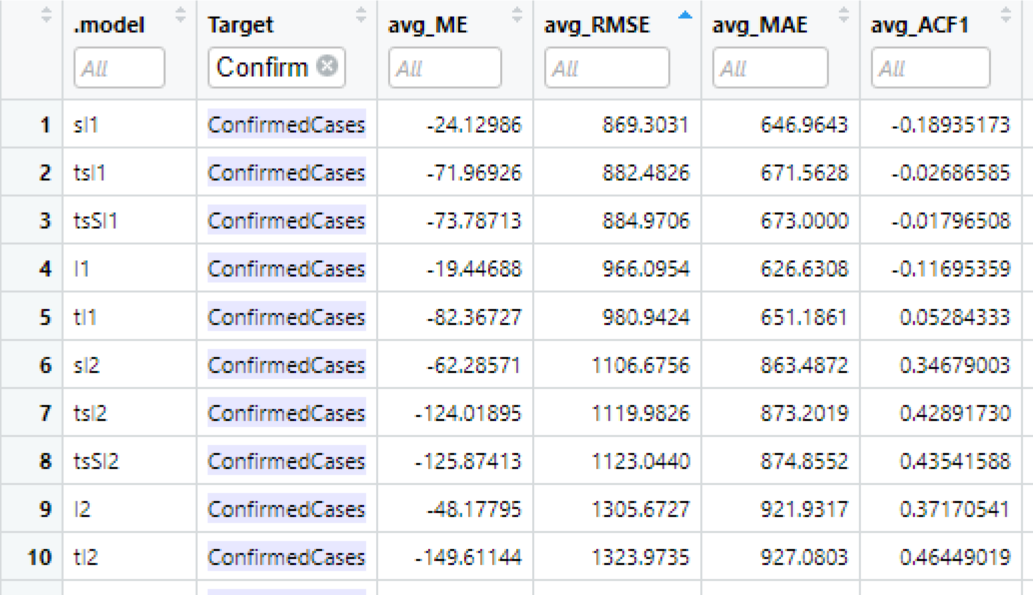
log(TargetValue) ~ Trend + season + PandemicStart + lag(log(TargetValue))

First, I decided to build the models using only the data for 4 countries, US, Canada, Italy and Germany, this is done because my computer is old and slow, so running the hole dataset will take a lot of time. The results are averaged by Fatalities and Confirmed Cases, because if we don't do it, it will give the results for each of the countries, and in return the results will be harder to compare. Models' names: t → trend, s → season, w → weight, S → PandemicStart, l<number> → lag, log → using log.



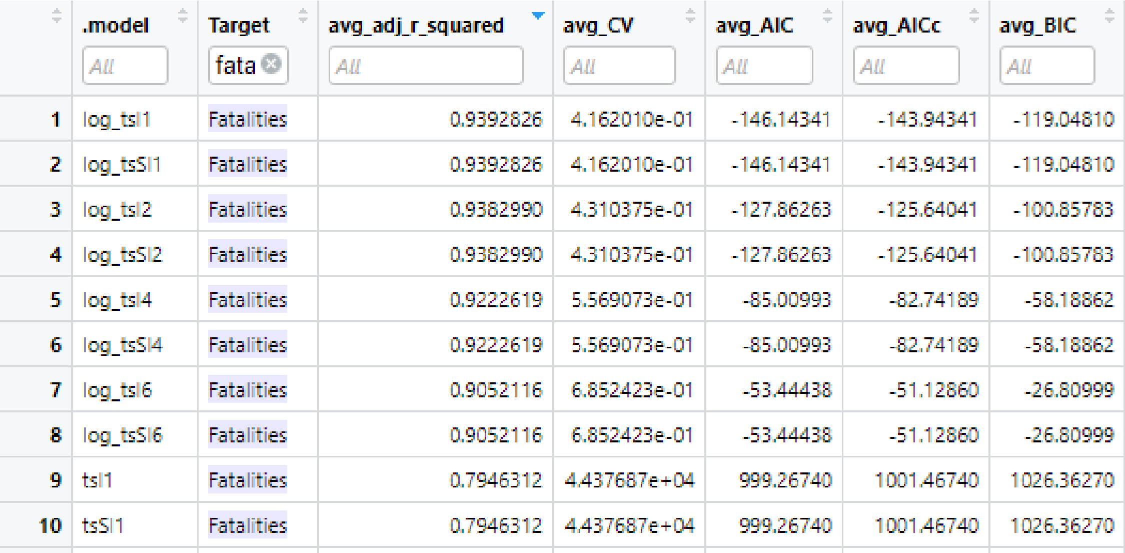
| | .model | Target | avg_ME | avg_RMSE | avg_MAE | avg_ACF1 |
|---|---|---|---|---|---|---|
| 1 | tsl1 | Fatalities | -28.525028 | 163.3275 | 98.73008 | -0.01366571 |
| 2 | tsSl1 | Fatalities | -28.525028 | 163.3275 | 98.73008 | -0.01366571 |
| 3 | sl1 | Fatalities | -10.018446 | 168.5459 | 90.91792 | -0.09996614 |
| 4 | tl1 | Fatalities | -30.647154 | 172.1925 | 98.51192 | 0.02185236 |
| 5 | l1 | Fatalities | -9.495386 | 177.0179 | 88.07641 | -0.07202595 |
| 6 | log_tsl1 | Fatalities | -64.212480 | 186.6867 | 98.72109 | 0.01834980 |
| 7 | log_tsSl1 | Fatalities | -64.212480 | 186.6867 | 98.72109 | 0.01834980 |
| 8 | tsl2 | Fatalities | -44.960059 | 212.0099 | 134.07896 | 0.45841095 |
| 9 | tsSl2 | Fatalities | -44.960059 | 212.0099 | 134.07896 | 0.45841095 |
| 10 | sl2 | Fatalities | -19.330207 | 219.2942 | 123.36310 | 0.42444220 |

*Figure 1 - First 10 models by Avg RMSE for fatalities (US, Canada, Italy, Germany)*



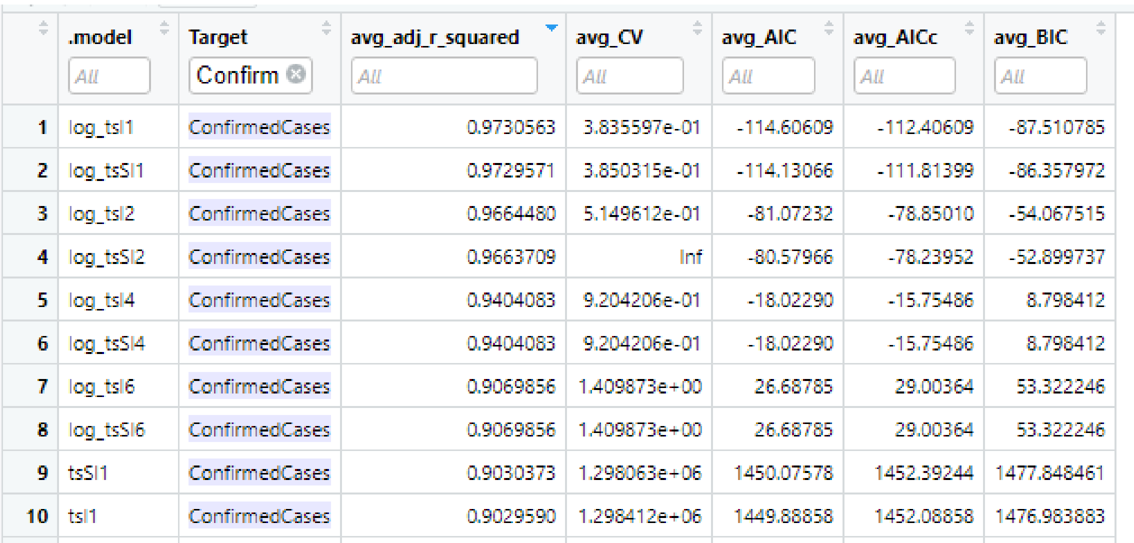
| | .model | Target | avg_ME | avg_RMSE | avg_MAE | avg_ACF1 |
|---|---|---|---|---|---|---|
| | All | Confirm ⊗ | All | All | All | All |
| 1 | sl1 | ConfirmedCases | -24.12986 | 869.3031 | 646.9643 | -0.18935173 |
| 2 | tsl1 | ConfirmedCases | -71.96926 | 882.4826 | 671.5628 | -0.02686585 |
| 3 | tsSl1 | ConfirmedCases | -73.78713 | 884.9706 | 673.0000 | -0.01796508 |
| 4 | l1 | ConfirmedCases | -19.44688 | 966.0954 | 626.6308 | -0.11695359 |
| 5 | tl1 | ConfirmedCases | -82.36727 | 980.9424 | 651.1861 | 0.05284333 |
| 6 | sl2 | ConfirmedCases | -62.28571 | 1106.6756 | 863.4872 | 0.34679003 |
| 7 | tsl2 | ConfirmedCases | -124.01895 | 1119.9826 | 873.2019 | 0.42891730 |
| 8 | tsSl2 | ConfirmedCases | -125.87413 | 1123.0440 | 874.8552 | 0.43541588 |
| 9 | l2 | ConfirmedCases | -48.17795 | 1305.6727 | 921.9317 | 0.37170541 |
| 10 | tl2 | ConfirmedCases | -149.61144 | 1323.9735 | 927.0803 | 0.46449019 |

*Figure 2 - First 10 models by Avg RMSE for Conformed Cases (US, Canada, Italy, Germany)*



| | .model | Target | avg_adj_r_squared | avg_CV | avg_AIC | avg_AICc | avg_BIC |
|---|---|---|---|---|---|---|---|
| | All | fata | All | All | All | All | All |
| 1 | log_tsl1 | Fatalities | 0.9392826 | 4.162010e-01 | -146.14341 | -143.94341 | -119.04810 |
| 2 | log_tsSl1 | Fatalities | 0.9392826 | 4.162010e-01 | -146.14341 | -143.94341 | -119.04810 |
| 3 | log_tsl2 | Fatalities | 0.9382990 | 4.310375e-01 | -127.86263 | -125.64041 | -100.85783 |
| 4 | log_tsSl2 | Fatalities | 0.9382990 | 4.310375e-01 | -127.86263 | -125.64041 | -100.85783 |
| 5 | log_tsl4 | Fatalities | 0.9222619 | 5.569073e-01 | -85.00993 | -82.74189 | -58.18862 |
| 6 | log_tsSl4 | Fatalities | 0.9222619 | 5.569073e-01 | -85.00993 | -82.74189 | -58.18862 |
| 7 | log_tsl6 | Fatalities | 0.9052116 | 6.852423e-01 | -53.44438 | -51.12860 | -26.80999 |
| 8 | log_tsSl6 | Fatalities | 0.9052116 | 6.852423e-01 | -53.44438 | -51.12860 | -26.80999 |
| 9 | tsl1 | Fatalities | 0.7946312 | 4.437687e+04 | 999.26740 | 1001.46740 | 1026.36270 |
| 10 | tsSl1 | Fatalities | 0.7946312 | 4.437687e+04 | 999.26740 | 1001.46740 | 1026.36270 |

*Figure 3 - First 10 models by Avg Adj. R$^2$ for fatalities (US, Canada, Italy, Germany)*



| | .model | Target | avg_adj_r_squared | avg_CV | avg_AIC | avg_AICc | avg_BIC |
|---|---|---|---|---|---|---|---|
| | All | Confirm | All | All | All | All | All |
| 1 | log_tsl1 | ConfirmedCases | 0.9730563 | 3.835597e-01 | -114.60609 | -112.40609 | -87.510785 |
| 2 | log_tsSl1 | ConfirmedCases | 0.9729571 | 3.850315e-01 | -114.13066 | -111.81399 | -86.357972 |
| 3 | log_tsl2 | ConfirmedCases | 0.9664480 | 5.149612e-01 | -81.07232 | -78.85010 | -54.067515 |
| 4 | log_tsSl2 | ConfirmedCases | 0.9663709 | Inf | -80.57966 | -78.23952 | -52.899737 |
| 5 | log_tsl4 | ConfirmedCases | 0.9404083 | 9.204206e-01 | -18.02290 | -15.75486 | 8.798412 |
| 6 | log_tsSl4 | ConfirmedCases | 0.9404083 | 9.204206e-01 | -18.02290 | -15.75486 | 8.798412 |
| 7 | log_tsl6 | ConfirmedCases | 0.9069856 | 1.409873e+00 | 26.68785 | 29.00364 | 53.322246 |
| 8 | log_tsSl6 | ConfirmedCases | 0.9069856 | 1.409873e+00 | 26.68785 | 29.00364 | 53.322246 |
| 9 | tsSl1 | ConfirmedCases | 0.9030373 | 1.298063e+06 | 1450.07578 | 1452.39244 | 1477.848461 |
| 10 | tsl1 | ConfirmedCases | 0.9029590 | 1.298412e+06 | 1449.88858 | 1452.08858 | 1476.983883 |

*Figure 4 - First 10 models by Avg Adj. R2 for Confirmed Cases (US, Canada, Italy, Germany)*

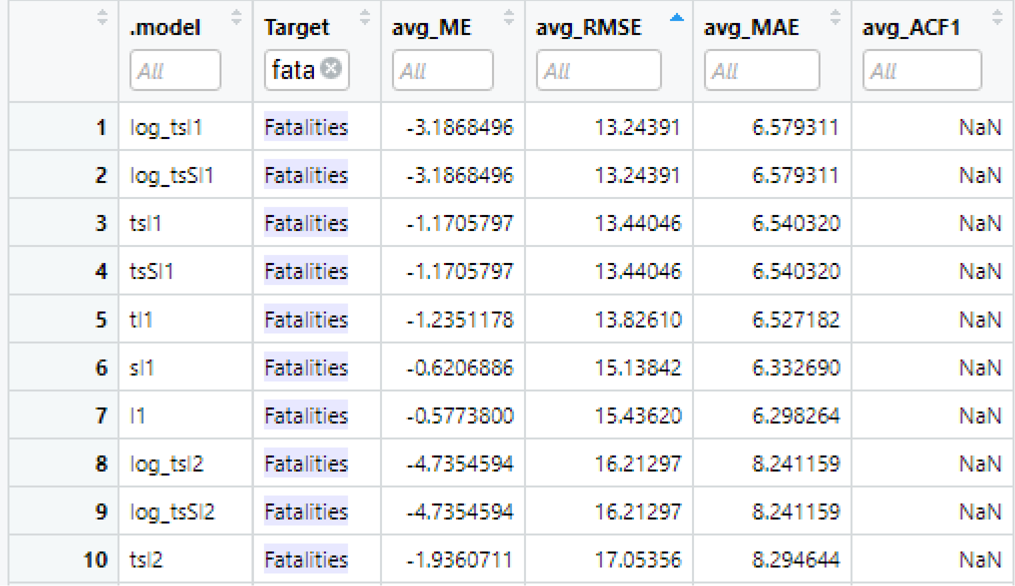
**Weight variable**: We can see that the models that used the variable weights as a predictor obtained the worst results, this is expected since this variable do not change in any scenario, and only changed between scenarios, thus resulting in not a good predictor.

**PandemicStart variable**: we can see that using or not using this variable in the models don't really make a big difference in the results, but overall, we can see that the models that used this variable underperformed with respect to their counterparts.

**Lag**: We can see that using lag increase the performance of the model, we can see that all the models with l<number lag> outperform the models without it, this means that the current value depends on previous ones, we also see that a the lower the lag is the greater performance we get, meaning that a lag of 1 achieved the best results. We can see that the best models in each of the tables uses a lag of 1.

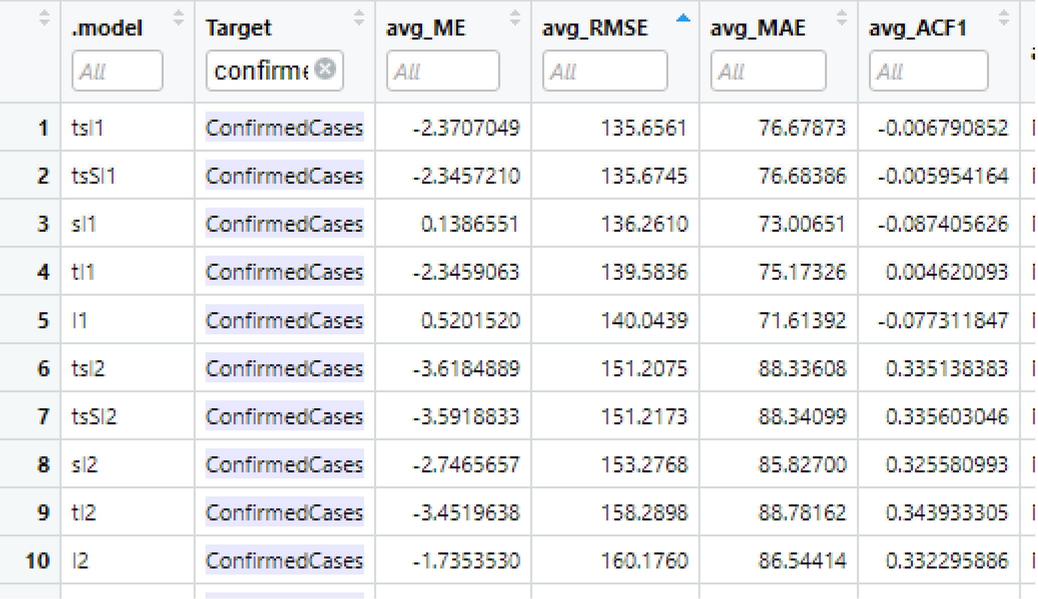
**Log**: We can see that the log models where able to outperform the rest of them if we rank them using the adj. R$^2$, achieving values of over 0.9. But if we look at the ranking using the RMSE, we can see that the log models no longer top the ranking.

Now we run the same models for the whole countries (We still don't use the provinces or counties), now we get the following results:



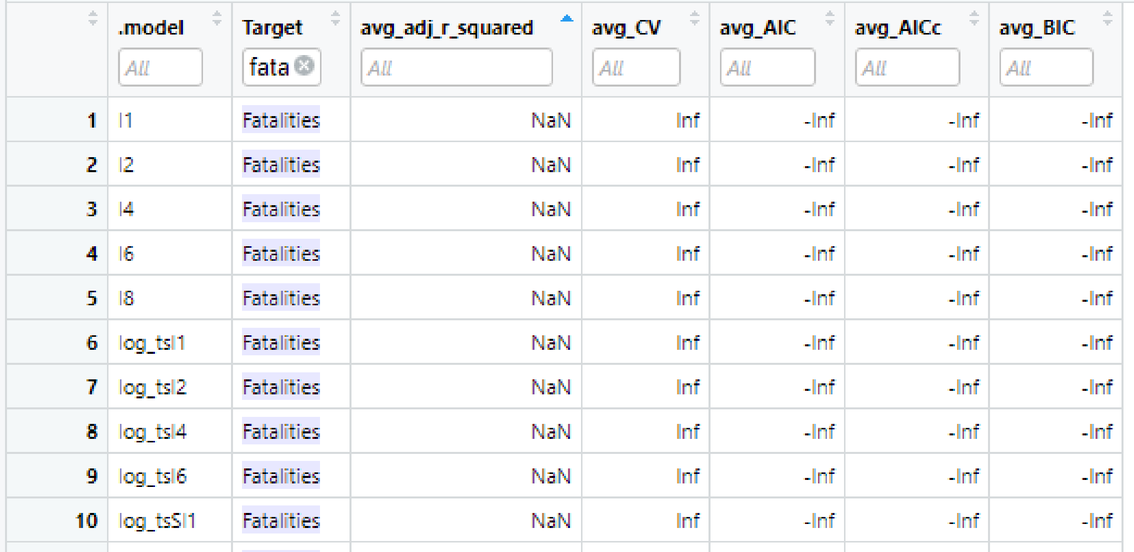
| | .model | Target | avg_ME | avg_RMSE | avg_MAE | avg_ACF1 |
|---|---|---|---|---|---|---|
| | All | fata | All | All | All | All |
| 1 | log_tsl1 | Fatalities | -3.1868496 | 13.24391 | 6.579311 | NaN |
| 2 | log_tsSl1 | Fatalities | -3.1868496 | 13.24391 | 6.579311 | NaN |
| 3 | tsl1 | Fatalities | -1.1705797 | 13.44046 | 6.540320 | NaN |
| 4 | tsSl1 | Fatalities | -1.1705797 | 13.44046 | 6.540320 | NaN |
| 5 | tl1 | Fatalities | -1.2351178 | 13.82610 | 6.527182 | NaN |
| 6 | sl1 | Fatalities | -0.6206886 | 15.13842 | 6.332690 | NaN |
| 7 | l1 | Fatalities | -0.5773800 | 15.43620 | 6.298264 | NaN |
| 8 | log_tsl2 | Fatalities | -4.7354594 | 16.21297 | 8.241159 | NaN |
| 9 | log_tsSl2 | Fatalities | -4.7354594 | 16.21297 | 8.241159 | NaN |
| 10 | tsl2 | Fatalities | -1.9360711 | 17.05356 | 8.294644 | NaN |

*Figure 5 - First 10 models by Avg RMSE for fatalities (all countries)*



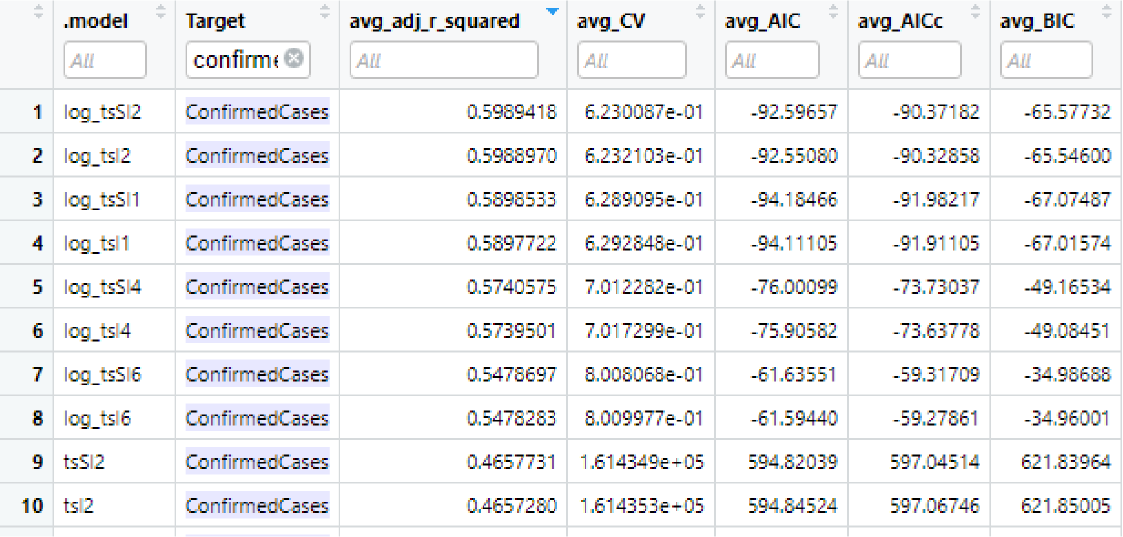
| | .model | Target | avg_ME | avg_RMSE | avg_MAE | avg_ACF1 |
|---|---|---|---|---|---|---|
| | All | confirme | All | All | All | All |
| 1 | tsl1 | ConfirmedCases | -2.3707049 | 135.6561 | 76.67873 | -0.006790852 |
| 2 | tsSl1 | ConfirmedCases | -2.3457210 | 135.6745 | 76.68386 | -0.005954164 |
| 3 | sl1 | ConfirmedCases | 0.1386551 | 136.2610 | 73.00651 | -0.087405626 |
| 4 | tl1 | ConfirmedCases | -2.3459063 | 139.5836 | 75.17326 | 0.004620093 |
| 5 | l1 | ConfirmedCases | 0.5201520 | 140.0439 | 71.61392 | -0.077311847 |
| 6 | tsl2 | ConfirmedCases | -3.6184889 | 151.2075 | 88.33608 | 0.335138383 |
| 7 | tsSl2 | ConfirmedCases | -3.5918833 | 151.2173 | 88.34099 | 0.335603046 |
| 8 | sl2 | ConfirmedCases | -2.7465657 | 153.2768 | 85.82700 | 0.325580993 |
| 9 | tl2 | ConfirmedCases | -3.4519638 | 158.2898 | 88.78162 | 0.343933305 |
| 10 | l2 | ConfirmedCases | -1.7353530 | 160.1760 | 86.54414 | 0.332295886 |

*Figure 6 - First 10 models by Avg RMSE for Conformed Cases (all countries)*



| | .model | Target | avg_adj_r_squared | avg_CV | avg_AIC | avg_AICc | avg_BIC |
|---|---|---|---|---|---|---|---|
| | All | fata | All | All | All | All | All |
| 1 | l1 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |
| 2 | l2 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |
| 3 | l4 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |
| 4 | l6 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |
| 5 | l8 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |
| 6 | log_tsl1 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |
| 7 | log_tsl2 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |
| 8 | log_tsl4 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |
| 9 | log_tsl6 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |
| 10 | log_tsSl1 | Fatalities | NaN | Inf | -Inf | -Inf | -Inf |

*Figure 7 - First 10 models by Avg Adj. R2 for fatalities (all countries)*



| | .model | Target | avg_adj_r_squared | avg_CV | avg_AIC | avg_AICc | avg_BIC |
|---|---|---|---|---|---|---|---|
| | All | confirme | All | All | All | All | All |
| 1 | log_tsSl2 | ConfirmedCases | 0.5989418 | 6.230087e-01 | -92.59657 | -90.37182 | -65.57732 |
| 2 | log_tsl2 | ConfirmedCases | 0.5988970 | 6.232103e-01 | -92.55080 | -90.32858 | -65.54600 |
| 3 | log_tsSl1 | ConfirmedCases | 0.5898533 | 6.289095e-01 | -94.18466 | -91.98217 | -67.07487 |
| 4 | log_tsl1 | ConfirmedCases | 0.5897722 | 6.292848e-01 | -94.11105 | -91.91105 | -67.01574 |
| 5 | log_tsSl4 | ConfirmedCases | 0.5740575 | 7.012282e-01 | -76.00099 | -73.73037 | -49.16534 |
| 6 | log_tsl4 | ConfirmedCases | 0.5739501 | 7.017299e-01 | -75.90582 | -73.63778 | -49.08451 |
| 7 | log_tsSl6 | ConfirmedCases | 0.5478697 | 8.008068e-01 | -61.63551 | -59.31709 | -34.98688 |
| 8 | log_tsl6 | ConfirmedCases | 0.5478283 | 8.009977e-01 | -61.59440 | -59.27861 | -34.96001 |
| 9 | tsSl2 | ConfirmedCases | 0.4657731 | 1.614349e+05 | 594.82039 | 597.04514 | 621.83964 |
| 10 | tsl2 | ConfirmedCases | 0.4657280 | 1.614353e+05 | 594.84524 | 597.06746 | 621.85005 |

*Figure 8 - - First 10 models by Avg Adj. R2 for Confirmed Cases (all countries)*

It appears that there was some error, since some of the outputs are NaN, this looks like it happens in the fatality's series, this could be due to a missing value or some weird anomaly, but since this part took so long to run, I preferred to run everything for the hole world, but using only 4 models, this way we can fixed faster the errors.

Now that we have the results of the hole countries, we need to build the model for the hole dataset, including provinces and counties, but since it took a lot of time to run the models on only the countries, in this case I'm only running one model:

TargetValue ~ Trend + season + lag(TargetValue, n=1)

And this are the results:

<PHOTOS>

(My computer is unable to run it, since now its 24:00 of the Sunday I am afraid I won't be able to show these results or make the predictions on the test set.)

## Limitations

In all of this models we haven’t used external datasets, this is short of a limitation since there are outside factors that affect the model, such as travel intensity, investments in healthcare, temperature and many other variables. Even factors that are extremely hard to categorize as variables, such as culture habits which is hard to represent in a numeric variable. Some cultures have a lot of person closeness like Mediterranean cultures, and other are the opposite like Scandinavian ones, and many other differences.

Another big difference is how does each country and region categories it’s diagnose cases and fatalities, some regions are going to make test to dead people to see if they had the virus, other countries won’t do it, also some countries won’t count the same way, as if a person dies form some other cause but tested positive some countries will add it to the covid statistics and others not.

Also, the number of diagnosed cases is not the same as the true infected population, this is because of not enough tests and asymptomatic people, this results in having a smaller sample of the actual infected population, but the new cases come from the actual infected population and not the diagnose one.

## Future Work

Future work can come from three different places. First, we can use more predictor variables in order to try to increase the performance of the models, such as the ones mentioned in the previous section.

Second, we can try to use more complex models such as neural networks and advance AI in order to build better models that are able to adapt to multiple waves and different covid strains.

Third, by trying to obtain better information by increasing testing and trying to standardize the procedures and systems different countries used. This will result in having better data and this will in turn result in better forecasting.

## What was learned/Conclusions

I learn how to work with big datasets in R and being able to construct models in order to make forecast or predictions into the future. This is a useful tool since this allows to get a glimpse into the future, by doing this we are able to prepare and plan accordingly, this means that we are able to reduce the impact of negative events and increase the impact of positive ones.

In this case the best models where the ones using trend season and lag, we have seen that using a lag of 1 is the best option.