Analysis of COVID-19 cases through Regression Models

Emilio Padrón Molina
emiliopm1997@gmail.com

October 10, 2020

Introduction

For many people, 2020 was a year that could only be possible in movies, bringing sorrow and concern in a worldwide scale. It is evident that this COVID-19 pandemic will be historically significant, which is why many researchers are starting to look for ways to prevent similar events from happening. One of these efforts could be to analyze the spread and characteristics that this virus had in every country, in other words, through data science.

In general, many people have tried to explain themselves how governments lost control over this crisis; nevertheless, this question is quite broad, and a more appropriate one could be: What are the specific characteristics that affected the way COVID-19 spread? An answer to this could be of great help to institutions such as the World Health Organization, and to international leaders given that they could learn from their and others' mistakes, and be better prepared for the next pandemic.

In this report, two linear regression models will be built as an attempt to answer the question in the previous paragraph. The first one will try to model the number of positive COVID-19 cases per million people that a country had for a month after having its first case of the virus; while the second one will attempt to model the maximum rate of change that a specific country had. The variables for these models will be the same (with the exception population of 2019), and will include GDP per capita, healthy life expectancy, among others.

Data

For this particular problem, data came from three different tables. The first one came from John Hopkins University, and it included the number of accumulated COVID-19 positive cases per day (from January 22nd to April 30th), per country; however, for many countries such as Canada and China, the accumulated cases were further divided by provinces or regions. The second data set was from the United Nations website and had information of many countries' GDP per capita, healthy life expectancy, social support, and others. The third data set reported every country/region's population of 2019 (before the pandemic had a global impact), and came from various sources including the UN, worldometers, statista,

and government websites.

The first step to build the final data set was to join the three different sources based on country names. This was done through the `INNER JOIN` instruction; this because all of the values in the final version of the set must be complete. However, not all of the values such as the longitude and latitude are relevant, therefore, these columns were deleted. Also, for countries that were divided by provinces and regions, both values were merged into a single column, that was then assigned as an index to have a more compact data set. After this initial cleanup, the data set contained 219 rows representing specific countries and regions, and 105 columns that included the accumulated cases per day, happiness score, GDP per capita, social support score, healthy life expectancy, perception of freedom to make life choices score, and population of 2019.

Methodology

The main focus at this stage was to obtain the target variables for the models. For both cases, the only relevant information was the accumulated positive cases per day. In the first model, an algorithm was built to detect the date were a given country had it's first case, then count thirty days, acquire the number of accumulated cases of that given day, and save it in a new column. Then, this number was divided by the country's corresponding population of 2019 and multiplied by a million to obtain the number of positive cases during the first month per country per million people; i.e. for a given country:

$$\text{Positive Cases in 1st Month per Million People} = \frac{\text{Total of cases in day 30}}{\text{Population 2019}} * 1000000$$

From all of the obtained results, it was relevant to analyze which countries had the highest values for this index. These results can be observed in Figure 1, which coincide with European-country dependent territories and small European countries. Additionally, before doing the model, the correlations between this target and the rest of the features were calculated as in Figure 2. It is evident that the variable that has the highest correlation with the number of positive cases in the 1st month of contagion per million people is the happiness score with 0.4. Even though this maximum correlation corresponds to a mild one it was decided that it was still worth attempting to make model.

On the other hand, for the second model, the maximum increase of positive cases in a day lapse was calculated for every country/region by using the accumulated cases per day data. In a similar manner as in the first model, it was relevant to analyze the countries that presented the highest increase in cases. The results can be observed in Figure 3 where it is clear that the countries and regions that appear are mainly developed countries, in the majority of cases from Europe. Moreover, the correlations between this target and the rest of the features were calculated as in Figure 4, for which the highest value was 0.136 corresponding to the healthy life expectancy variable. It is evident that this small correlation is also comparable to that of the GDP per capita and population of 2019; nevertheless, even

though these correlations aren't very significant, an attempt to build a predicting model with this data was still worth.

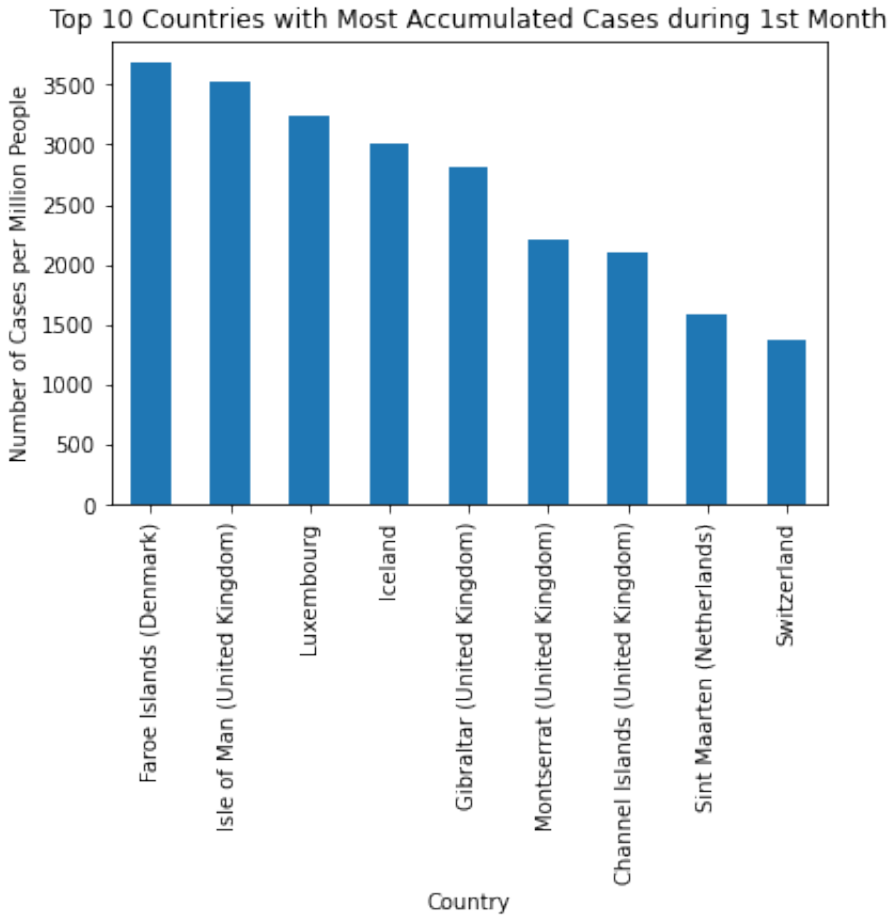

Figure 1: Top 10 countries/regions with highest accumulated cases during the first month from their initial case, per million people.

	Positive Cases per 1M in 1st Month of Infection
Happiness Score	0.407960
GDP per capita	0.345297
Social support	0.332634
Healthy life expectancy	0.304640
Freedom to make life choices	0.158251

Figure 2: Correlations between features and target variables on the 1st model.

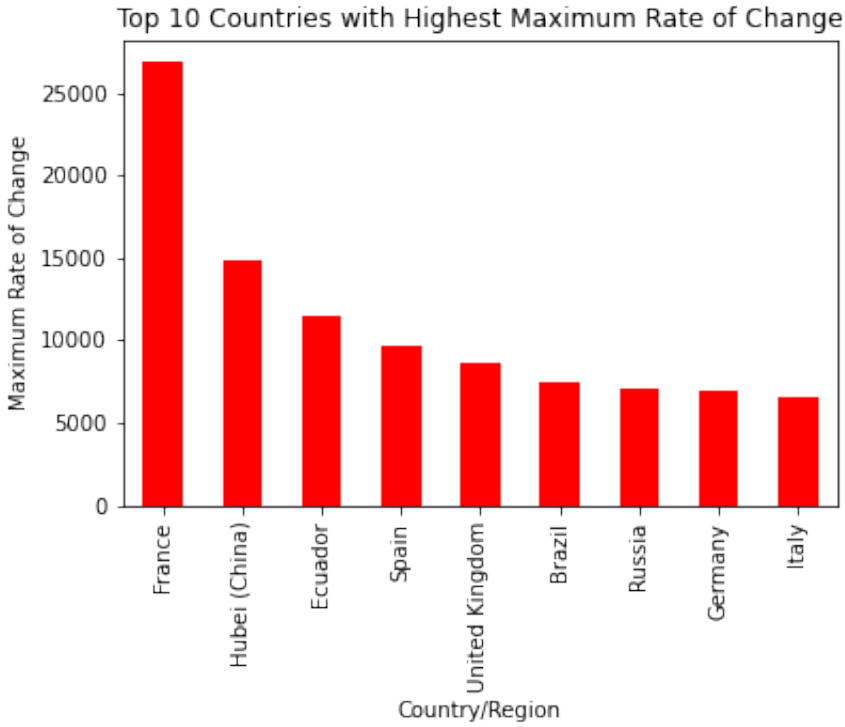

Figure 3: Top 10 countries/regions with highest maximum increase of positive COVID-19 cases in a day's lapse.

	Max Change
Happiness Score	0.104503
GDP per capita	0.131078
Social support	0.114900
Healthy life expectancy	0.136346
Freedom to make life choices	-0.027174
Population 2019	0.130437

Figure 4: Correlations between features and target variables on the 2nd model.

It is worth mentioning that for both models, countries and regions that had their first contagion in April were not considered for the correlations nor the model. This is because it was considered that more days of data are necessary to consider these instances as significant. Given this, the final data set was reduced to 211 rows and the feature columns (5 for the 1st and 6 for 2nd model respectively).

Results

Before generating a model, it is important to divide the data sets into features and target, and then once more into a training and testing splits, which in this case was done with a 85%-15% proportion. Then, the feature columns were scaled with the `MinMaxScaler` method from Scikit-learn.

The first model, whose goal was to predict the number of positive cases per 1M people in the 1st month of infection, was built using linear regression. This model resulted in the following weights for the features:

Features	Variable	Weights
Happiness Score	X1	1079.446162
GDP per capita	X2	329.674638
Social support	X3	-189.418783
Healthy life expectancy	X4	-223.068383
Freedom to make life choices	X5	-221.059989
Bias	1	-127.900372

Figure 5: Weights of the features in the first model.

Put into a mathematical equation, this means that the number of positive cases per 1M people in the 1st month of infection for a given country or region (NPC) can be calculated by:

$$\text{NPC} = 1079X_1 + 330X_2 - 189X_3 - 223X_4 - 221X_5 - 128 \tag{1}$$

In the second case, where the objective was to predict the maximum increase of cases in a country/region with specific characteristics, another linear regression model was constructed. This resulted in the coefficients:

Features	Variable	Weights
Happiness Score	X1	850.659221
GDP per capita	X2	-458.012612
Social support	X3	637.987848
Healthy life expectancy	X4	2703.569550
Freedom to make life choices	X5	-2440.175776
Population 2019	X6	6244.010876
Bias	1	-189.005366

Figure 6: Coefficients of the features in the second model.

According to the previous table, the equation to determine the maximum increase of cases

in a country/region (MIC) with its corresponding features is:

$$\text{MIC} = 850X_1 - 458X_2 + 637X_3 + 2703X_4 - 2440X_5 + 6244X_6 - 189 \tag{2}$$

Nonetheless, when these models were evaluated using their respective testing set, the obtained errors were quite large as it can be seen from Figure 7. Specially after looking at the R^2 score, one can notice that none of the models are accurate enough to be used for predictions. Reasons for this lack of prediction power will be discussed in the following section.

Model	MAE	MSE	R^2
NPC	248.72	237506	-3.16
MIC	677.12	653644	-0.62

Figure 7: Error measurements for NPC and MIC models.

Discussion

After analyzing the errors of the models in the previous section, it is evident that given that their R^2 values are negative, both models perform poorly when one uses them for predictions. As an attempt to fix this, it was relevant to generate regression plots of each target variable with their respective features. Examples of these are shown in Figures 8 & 9 where the feature with highest correlation in each model is plotted against the target.

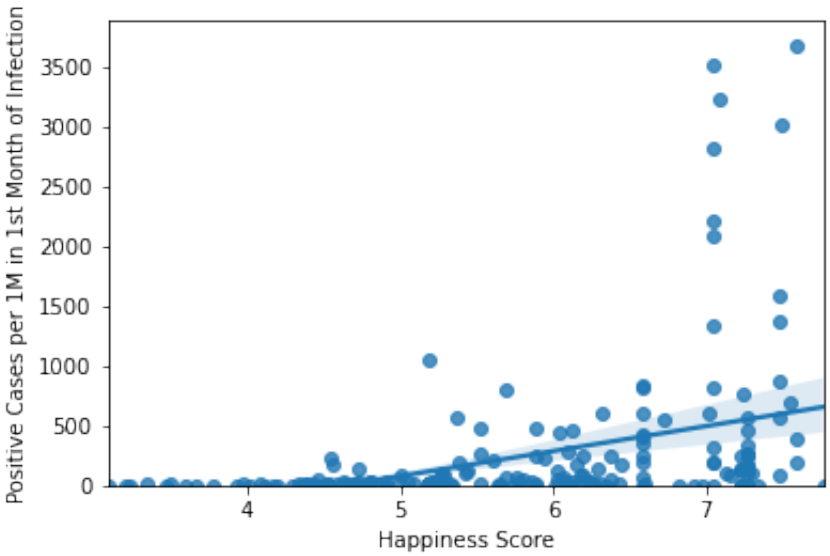

Figure 8: Regression plot of the number of positive cases per 1 million people during the 1st month of contagion with respect to the country's happiness score.

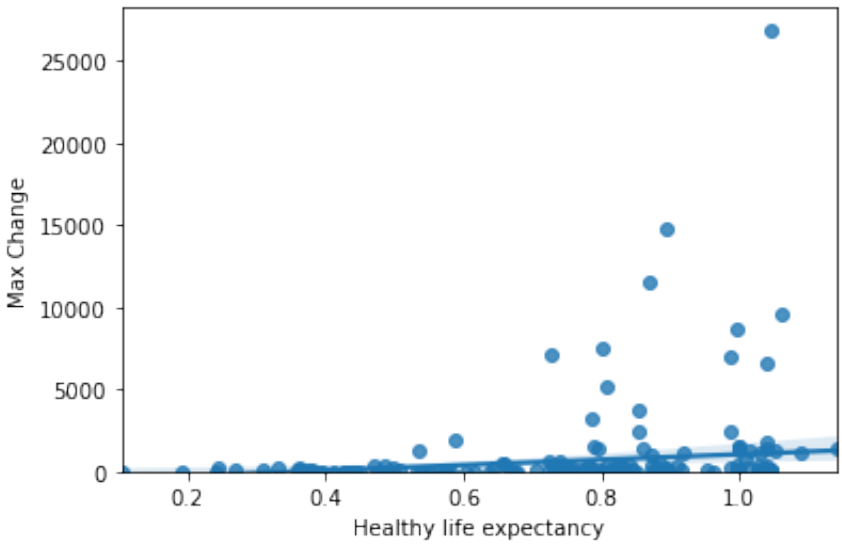

Figure 9: Regression plot of the maximum increase of cases in a day with respect to a country's healthy life expectancy.

The constructed plots portray to one of the main problems of the models, the relationships between the features and targets do not seem to be linear. Knowing this, both models were attempted to be rebuilt with pipelines such that polynomial features are taken into account. This would result in polynomial models of degrees two and three for each of the NPC and MIC sets. However, after testing each of these new models it was evident that even though the R^2 score increased, these were still less than zero, meaning that their performance was still not optimal.

After putting some thought, there are some actions that one can take to improve both models. If one was to use the same data sets, a more complex model could be built; nevertheless, this could result problematic given the amount of possibilities this could entail. One the other hand, one could also divide the data into more specific regions to have more rows to work with; although this might not make a remarkable improvement given the strength of the correlations. However, it is believed that in order to obtain more accurate models one should consider adding other features such as the date of quarantine of specific countries, other economic and health scores, and additional local actions taken by governments as a response to the pandemic. These new variables could result in stronger correlations and, as a consequence, better predictions.

Conclusion

In this report, three different data sets where used to model the total number of COVID-19 positive cases per 1M people in the first month of infection of a specific country and, also, the maximum daily increase of cases of countries. The three data sets included information from accumulated cases per country/region from January to April 2020, GDP per capita, happiness score, healthy life expectancy, population of 2019, and others. The most

important result from the analysis was that the happiness score is mildly correlated with the number of positive cases in the first month per 1M people with a value of 0.4. It is worth noting that the rest of the correlations for either model were small and therefore not very significant. Partly because of this, none of the linear models were very effective. Also, after generating a bar graph with the top 10 countries/regions with highest accumulated cases during the first month of their initial case, per million people, it was noticeable that these corresponded mainly to European-country dependent territories and smaller European countries. Moreover, a bar graph of the top 10 countries/regions with highest maximum increase of positive COVID-19 cases in a day's lapse was built. The countries that appeared in this figure corresponded with developed locations, mainly in Europe.

Additionally, it was noticed through regression plots, the feature variables and targets don't follow linear relationships, which is why polynomial models were built for improvement. Unfortunately even though these were better models, they were still not very accurate. Therefore, some recommendations were discussed to have a more effective model, the most noteworthy being to add more variables related to country's economic and health scores, and actions taken against the pandemic. Finally, it is worth noting that even though accurate models couldn't be built in this report, this exercise helped determine the importance of some variable in the world-wide spread of the virus.