Coping with covid: signs for success

Blog post at: https://datathisnthat.wordpress.com/2021/10/21/coping-with-covid-signs-for-success/

Diana Crowe, Eline Keutgens, Inkeri Virkki

1. Introduction

We analysed covid data for different countries as well as demographic and social data. We searched for patterns in the geodemographic data that can serve as indicators of how well or how poorly a country is capable of handling a crisis such as Covid-19. These markers would enable us to, in the case of future virus outbreaks, possibly predict which countries are going to get most affected by a virus.

Our data analysis showed us some unexpected results and allowed us to rule out several geodemographic variables. We have made our findings available online. Future similar studies including more variables and at a different stage of the pandemic could prove insightful.

2. Methods

2.1 Data Collection

We assembled the data for our analysis from several sources (see Appendix I: Data sources). We made sure to use only trustworthy sources to ensure data quality. The principal data was the total coronavirus infection and death figures for countries around the World made available by the John Hopkins Coronavirus Resource Center. We searched for and added extra columns for each country's: total population, area, population density, GDP, percentage of people aged 15 and over who can read and write, percentage of non-native inhabitants, percentage of economically active people who are also employed, and percentage of the population that is 65 years old and over.

2.2 Data Preparation or "pre-processing"

At this stage we cleaned and organized our raw data (data cleaning and wrangling):

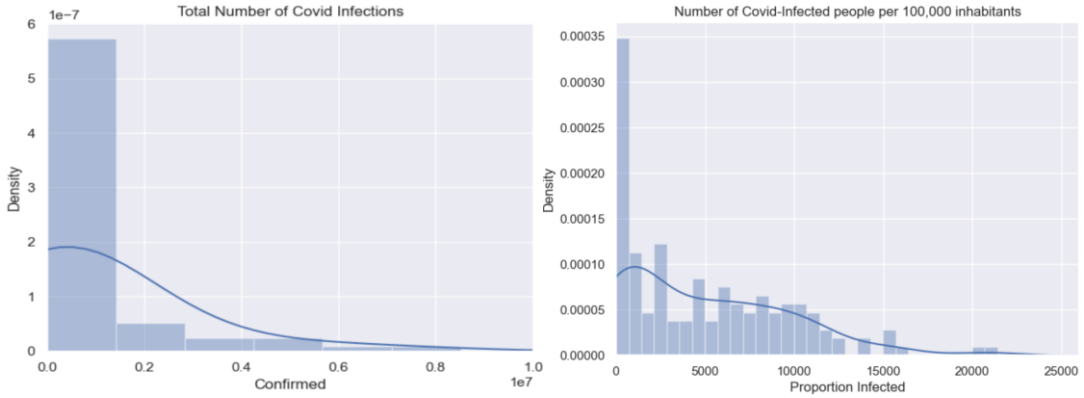
- We selected 149 of the countries available instead of using all of the countries listed and eliminated data rows containing "non-countries" (such as cruise ships).
- We eliminated superfluous data columns.
- We standardized the data for each country so that we could have total figures for the country instead of segregated by region.
- We calculated the population density of each country and added that as a new data column.
- We dealt with missing data: found and added missing country data from other reliable sources when possible; when no reliable values could be found, we calculated the mean and replaced all missing values with that value.

2.3 Data Processing

All of our data processing was done with Python, using algorithms to set the data in a format that is useful for further processing and machine learning.

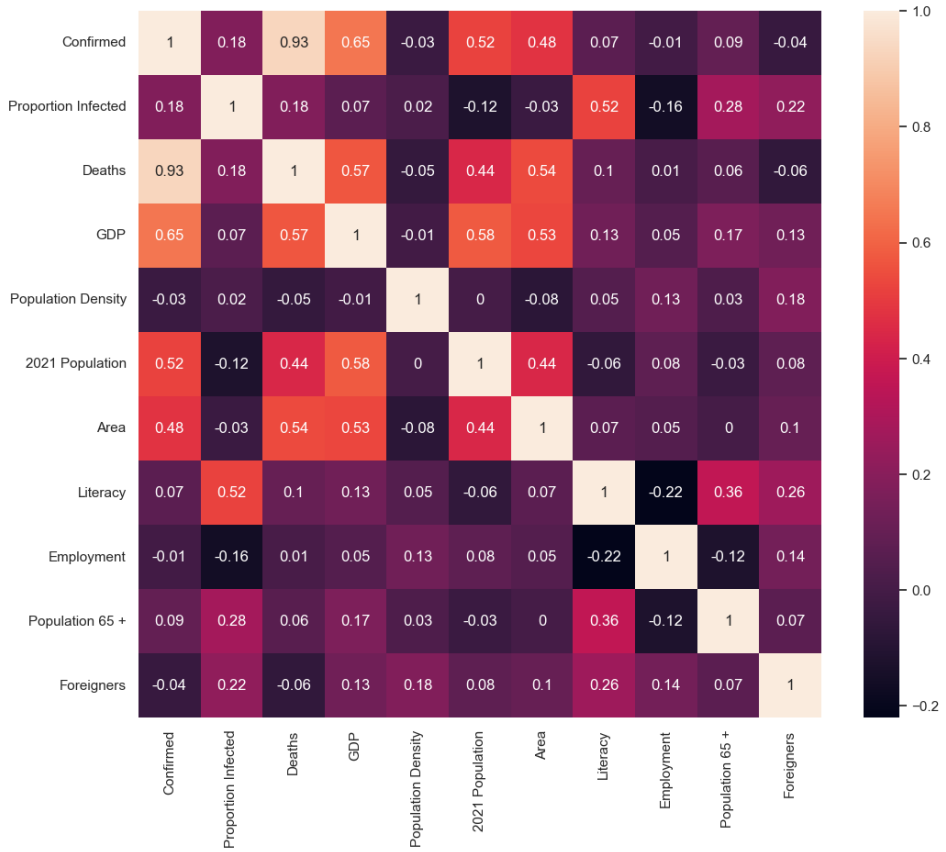
2.3.1 Exploratory Data Analysis

We looked at graphics of our covid-19 numbers and our geodemographic data. We also created correlation matrices using seaborn and a heatmap to search for correlations and find variables to use in regression models.



We see that this is not a Normal distribution… yet. It might still get to be a normal distribution in time. Currently there are still many countries with a small amount of people infected per 100,000 inhabitants and only a few countries with a really large amount of people infected per 100,000 inhabitants.

Correlation matrix using seaborn and a heatmap:



Observations:

- "Proportion Infected" and "Literacy" are moderately correlated.
- There is little if any linear correlation between "Proportion Infected" and any of the other variables…
- There is high correlation between the number of confirmed cases and the number of deaths
- The total number of confirmed covid cases has:
    - High correlation with the number of covid deaths
    - moderate correlation with the GDP
    - moderate correlation with the population of the country
    - low correlation with the area of the country
- The number of confirmed covid deaths has:
    - High correlation with the total number of covid cases
    - moderate correlation with the GDP
    - moderate correlation with the area of the country
    - low correlation with the population of the country
- There is moderate correlation between the GDP and the area of a country
- Although there is some correlation seen with the population and areas of a country, there is no correlation with the population density of a country.

2.3.2 Regression Analysis

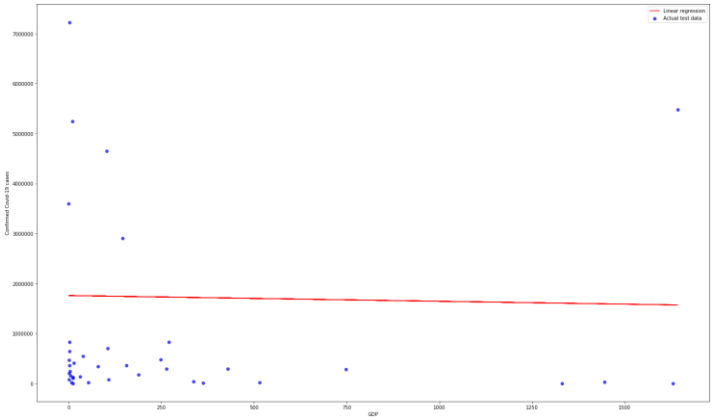
Logistic Regression:
We used Logistic Regression to predict whether a country would have a high or a low rate of covid cases. The barrier defined for getting into the high or low class is a proportion infected of 4000 (so that half of the countries would have a high infection rate and the other ones a low infection rate). As features we used population density, GDP, percentage of people aged 15 and over who can read and write, percentage of non-native inhabitants, percentage of economically active people who are also employed, and percentage of the population that is 65 years old and over. After splitting the data into a training and testing data (to get the best accuracy, we used about half of the data as training and the other half as testing data), we used the Logistic Regression module of the python library sklearn to fit the training. To make something interactive for the users to use, we made a widget where you change the values of the features to see whether your country would most likely have a high or low infection rate with those features.

Linear Regression:
Confirmed cases and GDP and 2021 Population:
From the correlation heatmap we noticed that the correlation between confirmed cases and GDP was 0,65 and the correlation between confirmed cases and population in 2021 was 0,52 (i.e. moderate correlations). We used linear regression to see if we can predict confirmed covid cases from GDP and total population. Linear Regression between confirmed cases and GDP and population had to be done separately because GDP and population also had a moderate correlation of 0,58, and with multiple linear regression it would be hard to specify the effect of just GDP or just the population.

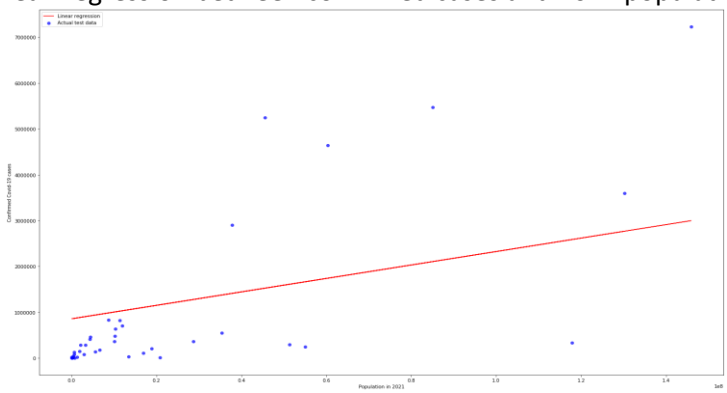
First, we read the csv-file where we had all our data into a pandas data-frame. After that we could easily split the columns "Confirmed", "GDP" and "2021 Population" from the data-frame to separate vectors. We used python's library sklearn and its linear model module to perform Linear Regression. We used the train_test_split function from the sklearn.model_selection module to split our data into training and testing

data. We then created a new Linear Regression model, fitted the training data in the model, computed the prediction with the sklearn linear regression predict-function and the test values. After that we plotted the Linear Regression and computed the score of the model (r-squared score) with the sklearn Linear Regression score-function.

Linear Regression between confirmed cases and GDP

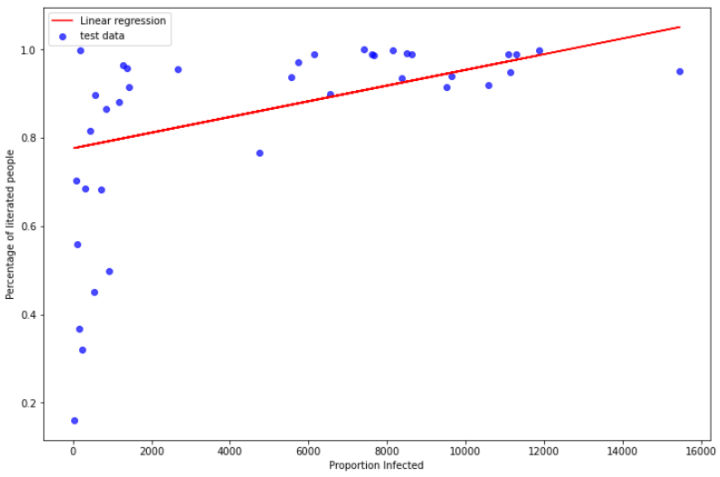


Linear Regression between confirmed cases and 2021 population



Proportion Infected and Literacy:

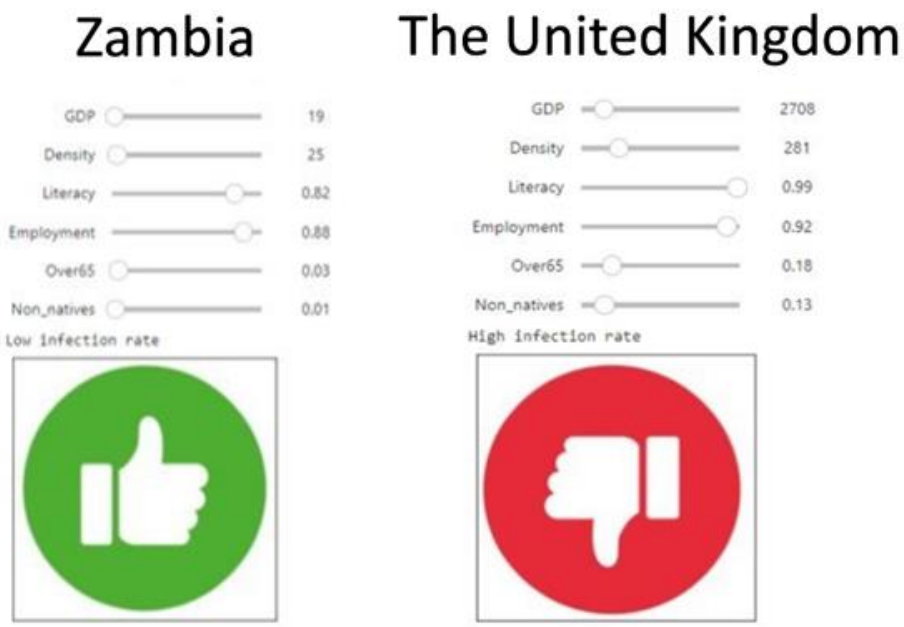
Similar to what was done for the correlation between the confirmed cases and GDP/population in 2021, we also made a linear regression model for the correlation between the proportion of infected people and the literacy. This is because we noticed in the correlation matrix that there is a moderate correlation between these features.



3. Results

Logistic Regression

For the Logistic Regression, the accuracy is about 72% which is not too high. We tried many different ways of improving the accuracy (by using another set of features, splitting the data into training and testing data differently, …) and this was the highest accuracy that we could get. On the other hand, considering that we are working with real data of real-life situations with much more features than we used, this is a decent accuracy. If you play with the widget created, you see that the percentage of people aged 15 and over who can read and write is the variable that has the highest weight in predicting if a country would have a high or a low infection rate. An example of a country that is correctly categorised as a 'good' country and one that is correctly categorised as a 'bad' country is given below.



Linear Regression

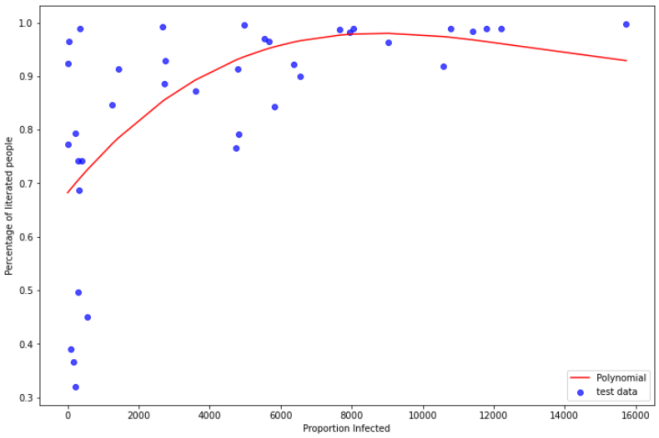
Confirmed cases and GDP and 2021 Population:

The R-squared score (r2 score) for the Linear Regression model between confirmed cases and GDP was -0,18. (The score varied a little bit with each run, but was always negative.) This means that the Linear Regression model didn't fit the trend of the data. We noticed from plotting 'Confirmed' and 'GDP' that we had a couple of outstanding datapoints. GDP in China and USA were much higher than in other countries, and USA, India and Brazil had huge numbers in confirmed cases. We tried the Linear Regression model without those peaking countries to see if we could get a better r2 score. This way we got the score to -0.006, but with each run it was still negative.

Linear Regression between confirmed cases and population in 2021 showed better results on the other hand. For the r2 score we got 0,33, which means that population can indicate the number of confirmed cases. We also tried the Linear Regression without Brazil, China, India and USA, but that actually gave worse results (0,23).

Proportion infected and literacy:

After doing a simple linear regression on this model, we found that the R2 score for the testing set is around 0.21, what is not really high. When plotting the data, it looked that if we would try a polynomial regression instead of just fitting a linear model, we might find a better fit. So, after trying polynomial regression with different degrees, we found that for a degree of 3, we got a R2 score of 0.4 which is already higher. We tried higher degrees but then we would overfit the data.



4. Conclusions

The country features that we looked at were not the "signs for success" that we were hoping to find. They were not significant in predicting how well a country would handle a situation like Covid-19. Only the percentage of people aged 15 and over who can read and write has a remarkable influence on the proportion of people getting infected. That is something that we not only see in the correlation matrix made with seaborn, but is also present in the Logistic Regression and the Linear Regression. The fact that there is an influence of the literacy on the proportion infected is not a causation. There is a moderate correlation between these two features but that does not mean that if you would change the literacy in your country because you were a so called 'bad' country, you would then end up with a 'good' country. What is also remarkable is that for that percentage of people aged 15 and over who can read and write, it is more likely to end up with a high infection rate if you have a higher percentage. This is visible in the examples of the widget shown above (Zambia has a pretty low literacy percentage while the United Kingdom has a really high literacy percentage). This means that the higher the literacy, the higher the infection rate in your country. Though we surprisingly found out that literacy has a strong effect on Covid cases, we believe that in this case correlation does not equal causation. We thought of other possible explanations for this. Countries with the highest literacy are likely to be the most developed ones. This means advanced healthcare available throughout the country that can provide enough places and tests for people to go get tested without them feeling it is a nuisance. The more developed countries can also have a better social security which allows people to take sick leave from work and to be able to return to work. Literate people can also have a better awareness of the Covid-19 situation in their country as well as the symptoms. This all leads to better and more accurate documentation of occurred cases of Covid. In addition, traveling most likely also plays a part in why Covid cases are higher in more literate countries. Traveling (and especially traveling abroad) is more common in more developed countries. Traveling between countries where literacy is high makes the virus spread in these countries. So the correlation between literacy versus proportion infected is most likely a case of confounding (Simpson's paradox): there's overall more cases in areas with more travel and denser, more urban population in areas where literacy is higher.

For the other features, we didn't find good correlations with the proportion infected. We clearly see out of this project that predicting if a country would handle a crisis like Covid-19 relies on so much more than the features that we used. For possible future projects, we could use the restrictions used during the pandemic, degree of mobility, respect for authority, … to make better models on how well a country would do and what the signs of success would be.

5. Our Experience

It was challenging to work with real data and at the same time try to get good results with the machine learning tools that we had. Also, the results in correlations between features and the proportion infected

were disappointing. We expected to see stronger, more useful correlations between the features that we used.
For putting the application (the widget in the notebook) online, we had several difficulties with Heroku. We first used Voilà to put the widget in a more beautiful form so that it is easy for the user to understand and to work with. But putting it online with Heroku within the time limit was not possible due to colliding dependency problems between Heroku and Jupyter Notebook. Therefore, we chose to deploy our Jupyter Notebook file with Github and Binder so that the notebook opens in your browser without the need of installing any program that can run a notebook (it is also possible to open and use it on your smartphone). This is not as beautiful as working with Heroku but we tried to make it as simple and user friendly as possible so that the main purpose of widget would not get lost.
(Binder link for the widget: https://mybinder.org/v2/gh/inkeriV/Coping-with-Covid-19-application/main?labpath=Coping_with_covid_19_signs_of_success_application.ipynb)

Appendix I: Data sources

1. Collected COVID19 data from the John Hopkins Coronavirus Resource Center: https://coronavirus.jhu.edu/about/how-to-use-our-data?fbclid=IwAR34LrKBj1-Ic4kT9F89Gu6r1oOQdBEriqa7PYlSPGVCcWKizH_Ezk90OMo

2. Collected Geodemographic and Social data from the United Nations: https://unstats.un.org/unsd/demographic-social/products/dyb/index.cshtml/?fbclid=IwAR277ZK_PLqKARFt8y9o-mDeoknwnq44lFJcbiPPVzQhUgau7R8Uqt7uYBo#censusdatasets

3. Gross domestic product (GDP): https://tradingeconomics.com/country-list/gdp

4. Country population sizes and densities: https://worldpopulationreview.com/country-rankings/countries-by-density and https://github.com/mledoze/countries