COVID-19 correlation with BCG vaccination policies a CRISP-DM approach

Diaz-Moraila J.E.
School of Sciences and Engineering
Tec de Monterrey
Monterrey, Mexico
A00828174@itesm.mx

***Abstract*—In December 2019, COVID-19 coronavirus was first identified in the Wuhan region of China. By March 11, 2020, the World Health Organization (WHO) categorized the COVID-19 outbreak as a pandemic. A lot has happened in the months in between with major outbreaks in Iran, South Korea, and Italy.**

We know that COVID-19 spreads through respiratory droplets, such as through coughing, sneezing, or speaking. But, how quickly did the virus spread across the globe? And whats the impact of the disease in different countries?.

Modeling the COVID-19 to answer to this questions is not something that's simple to accomplish, since there are different factors across countries that make this estimations fuzzy.

In this work its proposed a CRSIP-DM approach to explain one the factor that may lead to create a good model which explains national differences in COVID19 enriching data with different national policies respect to Bacillus Calmette-Guérin (BCG) childhood vaccination.

***Index Terms*—COVID-19, Modeling, BCG, vaccination, CRISP-DM**

I. Introduction

The COVID-19 pandemic originated in December 2019 in Wuhan China and it has quickly spread over all continents affecting most countries in the world. However, there are some striking differences on how COVID-19 is behaving in different countries.

Despite organizations around the world have been collecting data so that governments can monitor and learn from this pandemic. Notably, the Johns Hopkins University Center for Systems Science and Engineering created a publicly available data repository to consolidate this data from sources like the WHO, the Centers for Disease Control and Prevention (CDC), and the Ministry of Health from multiple countries, there exist gaps in several data entries that may lead to a good model that explain the COVID-19 behaviour.

A. Business understanding

As it is explained in **fivethirtyeigh** post [4] *”It's relatively easy to put together — the sort of thing people on our staff do while buzzed on a socially isolated conference call after work. The number of people who will die is a function of how many people could become infected, how the virus spreads and how many people the virus is capable of killing”.*

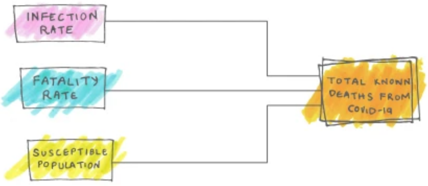


Fig. 1. COVID-19 model deconstruction archetype taken from [4]

”But then you start trying to fill in the blanks. Every variable is dependent on a number of choices and knowledge gaps. And if every individual piece of a model is wobbly, then the model is going to have as much trouble standing on its own as a data journalist who has spent too long on a conference call while socially isolated after work”.

It is known that Different countries and regions collect data in different ways. There's no single spreadsheet everyone is filling out that can easily allow us to compare cases and deaths around the world. Even within the United States, doctors say we're under-reporting the total number of deaths due to COVID-19 [6].

The same inconsistencies apply to who gets tested. Some countries are giving tests to anyone who wants one. Others are not [3]. That affects how much we can know about how many people have actually contracted COVID-19, versus how many people have tested positive.

And the virus itself is an unpredictable contagion, hurting some groups more than others [1], meaning that local demographics and health care access are going to be big determinants when it comes to the virus' impact on communities. Some partial explanation to predict this so called unpredictable phenomena is proposed in [5] where it is hypothesized that BCG vaccination policies can lead to explain this phenomena, since BCG is used widely across the world as a vaccine for Tuberculosis (TB), with many nations, including Japan and China, having a universal BCG vaccination policy in newborns 1. Other countries such as Spain, France, and Switzerland, have discontinued their universal vaccine policies due to comparatively low risk for developing M. bovis infections as well as the proven variable effectiveness in preventing adult TB; countries such as the United States, Italy, and the Netherlands, have yet to adopt universal vaccine policies for similar reasons.

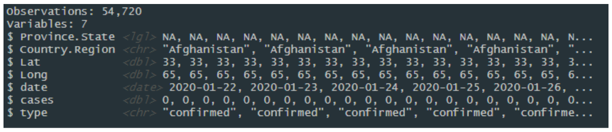
The main objective of this work is to provide some insights and further exploration of the correlation between countries that have BCG vaccination policies with the morbidity and mortality for COVID-19, as well as proposing some predictive models which can be trained with the same features; The whole process is done under the light of the Cross-industry Standard Process for Data Mining (CRISP-DM).

II. Materials and Methods

The COVID-19 dataset is obtained by the R package *coronavirus* where the raw data is pulled and arranged by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from the resources described in https://ramikrispin.github.io/coronavirus/#data-sources .

For the enriching part it is used the Immunization coverage dataset https://www.who.int/immunization/monitoring_surveillance/data/en/ and http://www.bcgatlas.org/.

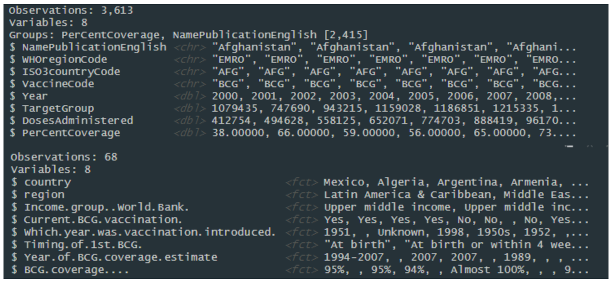
Coronavirus dataset is described by 7 features and 54,720 observations some of these features describe the case types (confirmed, death, recover), the date of this measurement and its quantity, next in figure 2 is a glimpse of all these features and observations.



```
Observations: 54,720
Variables: 7
$ Province.State <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
$ Country.Region <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", "...
$ Lat            <dbl> 33, 33, 33, 33, 33, 33, 33, 33, 33, 33, 33, 33, 33, 33, 33, 3...
$ Long           <dbl> 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 6...
$ date           <date> 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-25, 2020-01-26, ...
$ cases          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ type           <chr> "confirmed", "confirmed", "confirmed", "confirmed", "confirme...
```

Fig. 2. glimpse of coronavirus dataset

As well the immunization dataset and BCG atlas dataset that will be used to enrich this dataset in order to propose an alternative explanation: that the difference country by country in the different types of cases (confirmed, death, recovered) of COVID-19 can be partially explained by the national policies on vaccination against Bacillus Calmette-Guérin (BCG), a glimpse of both datasets is depicted as follows:



```
Observations: 3,613
Variables: 8
Groups: PerCentCoverage, NamePublicationEnglish [2,415]
$ NamePublicationEnglish <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghani...
$ WHOregionCode          <chr> "EMRO", "EMRO", "EMRO", "EMRO", "EMRO", "EMRO", "EMRO...
$ ISO3countryCode        <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG...
$ VaccineCode            <chr> "BCG", "BCG", "BCG", "BCG", "BCG", "BCG", "BCG", "BCG...
$ Year                   <dbl> 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008,...
$ TargetGroup            <dbl> 1079435, 747690, 943215, 1159028, 1186851, 1215335, 1...
$ DosesAdministered      <dbl> 412754, 494628, 558125, 652071, 774703, 888419, 96170...
$ PerCentCoverage        <dbl> 38.00000, 66.00000, 59.00000, 56.00000, 65.00000, 73....

Observations: 68
Variables: 8
$ country                                <fct> Mexico, Algeria, Argentina, Armenia, ...
$ region                                 <fct> Latin America & Caribbean, Middle Eas...
$ Income.group..World.Bank.              <fct> Upper middle income, Upper middle inc...
$ Current.BCG.vaccination.               <fct> Yes, Yes, Yes, Yes, No, No, , No, Yes...
$ Which.year.was.vaccination.introduced. <fct> 1951, , Unknown, 1998, 1950s, 1952, ,...
$ Timing.of.1st.BCG.                     <fct> "At birth", "At birth or within 4 wee...
$ Year.of.BCG.coverage.estimate          <fct> 1994-2007, , 2007, 2007, , 1989, , , ...
$ BCG.coverage....                       <fct> 95%, , 95%, 94%, , Almost 100%, , , 9...
```

Fig. 3. glimpse of WHO immunization and BCG Atlas dataset respectively

In this work as it was mentioned the CRISP-DM methodology is going to be used which is the most widely used form of data-mining model and the case study is no exception. The popularity of CRISP-DM comes from its various advantages for solving the existing problems related to data mining even across different industries thanks to its robustness. CRISP-DM breaks the process of data mining into six major phases as is shown in table 4 :

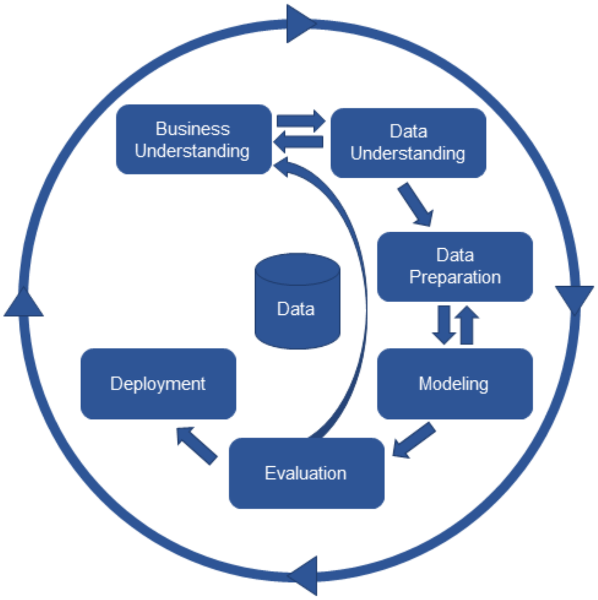


Fig. 4. CRISP-DM Methodology

A. Data understanding and preparation

To begin exploratory analysis, it could be informative to find out the mean and variance of the case types in the data, there are 3 possible case types : **confirmed, recovered, death** it is worth mention that for purposes of avoiding outliers and to much sparcity only the countries with 500+ confirmed cases where filtered, take in consideration that this filtering is performed on April 2nd of 2020.

Figure 5 shows a box plot for each type for all (500+ confirmed cases) countries as we can observe there still a wide range of values besides the filtering as shown in I

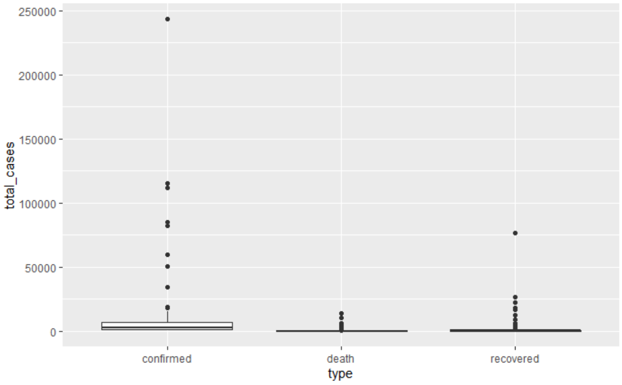


Fig. 5. Box plot of the different case types of COVID-19.

To deal with sparcity we split the data into 3 groups w we pick the 1/3 most extremist (figure 6) and (figure 8) confirmed

TABLE I
SUMMARY OF CONFIRMED CASES WORLDWIDE

Min:	505
1st Qu.:	1021
Median :	2454
Mean:	14706
3rd Qu.:	7154
Max.:	243453

cases and 1/3 from the middle (figure 8) from each type, after this grouping we can depict our data better.

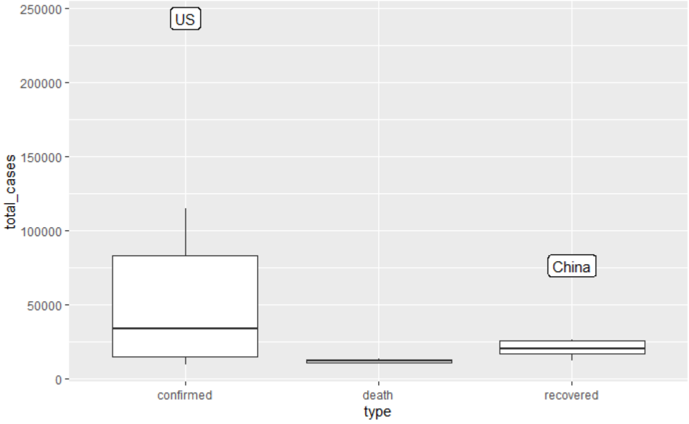


Fig. 6. Boxplot of group 1

It is depicted that US is clearly an outlier and the confirmed cases oscillate between 9976 and 243453 with a mean of 46939 and a standard deviation of 63257.77. on the other hand we have comparatively low cases of death cases with a mean of 12132 and a standard deviation of 2522, finally we can observe that recovered cases has a similar mean that death cases 28881 but has more variability with 23862.61 standard deviation.

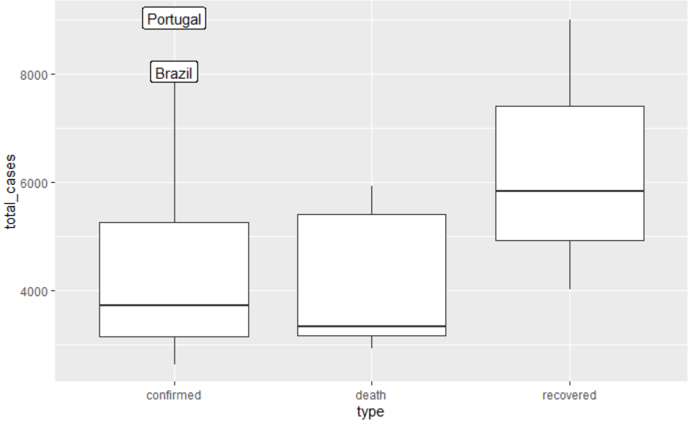


Fig. 7. Boxplot of group 2

In the second and third boxplot as expected we have smaller thresholds but more variability with, we can observe that figure 7 have more deaths an recovery cases relatively that the figure 6, and we can conlcude that indeed there are different behaviours between this groups not only in the number of confirmed cases but in their respective death and recovery rate.

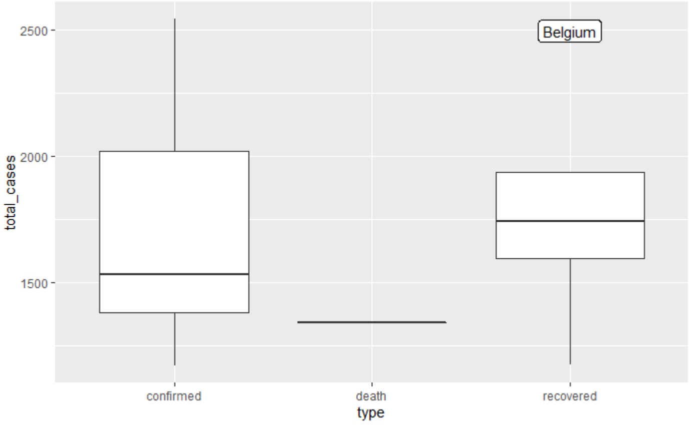


Fig. 8. Boxplot of group 3

Another exploratory question to ask about this dataset is if whether there is a significant difference in the means exhibited across different income or not.As an example, all observations for the 3 income groups according to World Bank (Figure 10) were taken and a one-way analysis of variance (one-way ANOVA) was performed .

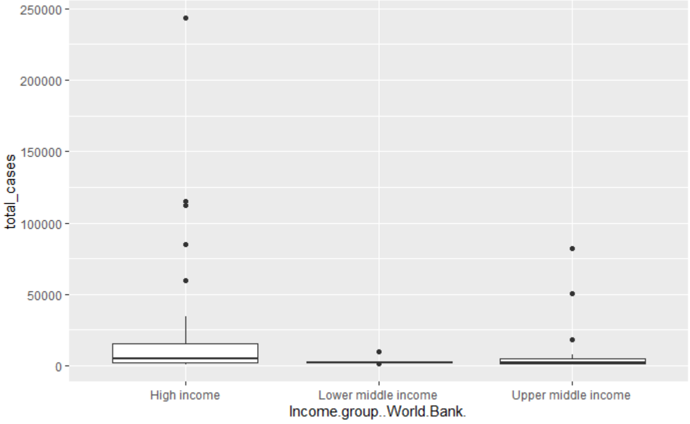
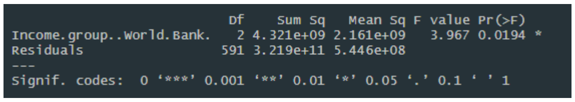


Fig. 9. boxplot of confirmed cases across income groups.

The summary statistics for the one-way ANOVA can be seen in Figure 10 it is observed that the F-value obtained in the test (3.9) is likely not due to chance (p-value 0.019 for a random Fischer distribution with 2 degrees of freedom), and so it can be confirmed there is a partially significant difference between the mean values of these features.



```
                           Df    Sum Sq   Mean Sq F value Pr(>F)  
Income.group..World.Bank.   2 4.321e+09 2.161e+09   3.967 0.0194 *
Residuals                 591 3.219e+11 5.446e+08                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 10. One-way ANOVA summary statistics for 3 income groups in Enriched coronavirus dataset.

B. Modeling for the data understanding

One of the applications of statistics is to extract inferences in populations from the study of samples. This process is called Inferential Statistics and its studies aim to deduce (infer) properties or characteristics of a population from a representative sample [2].
One of this tools is correlation and determines whether causal or not, between two random variables or bivariate data are dependent. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related.

In Figure 12 , a Pearson correlation heatmap of these scores is presented. The correlation between all scores is in general weak , however this may me due that just *total cases* is continuous, and the rest are categorical features (with a small range). Also what is important here, is that BCG vaccination catch a correlation trend between *total cases* although negative, (because value 1 is for NO and value 2 is for YES), meaning that if a country does not have a currently a national policies on Bacillus Calmette-Guérin (BCG) vaccination.

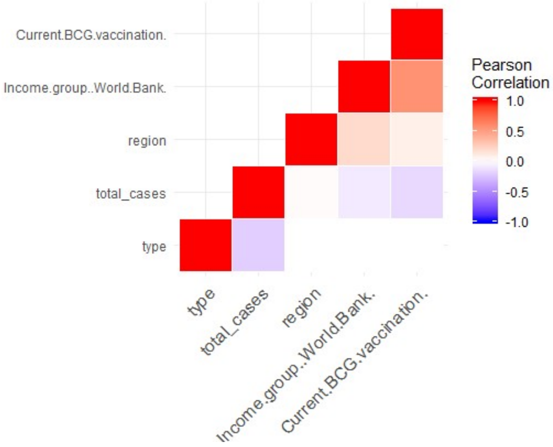


Fig. 11. Pearson correlations between all variables.

If we split the total cases by each type (Confirmed, Death, Recovered) we can observe more specifically that confirmed and death cases catch a weak negative trend meaning that countries that do not have currently a BCG vaccination policy expect to have more confirmed and death cases.

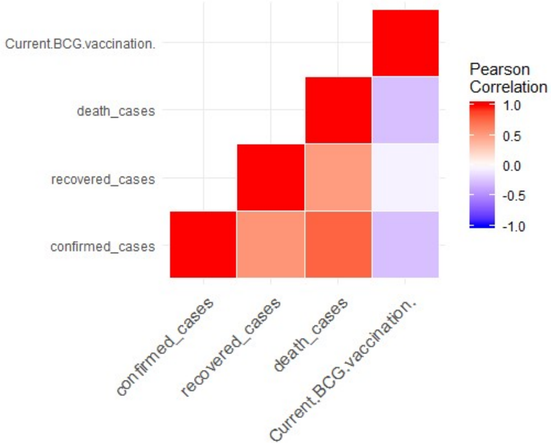
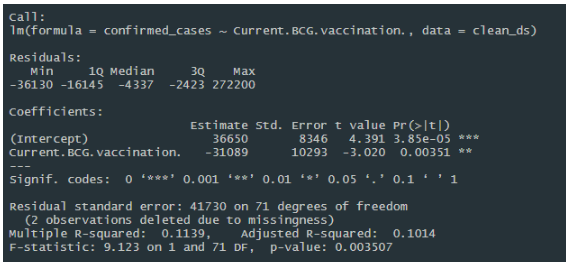


Fig. 12. Pearson correlations between BCG Vaccination policies .

The business goal, or the problem we are trying to solve through our data analysis is whether a country is prone to have high confirmed cases of COVID-19 through the correlation analysis with universal policies of BCG vaccination.

Now that we have defined *Current.BCG.Vaccination* as a variable associated with the predicted value *Confirmed_cases*, we can take a look at how the model fits the data. In our case, turns out the model poorly fits the data , having a big residual standard error (41730), a low value of r-squared and a low F-value. For our model we have 71 degrees of freedom, which can be considered quite high. This supports the evidence of rejecting the null hypothesis. This value ranges from zero to one, where a high value means most of the information can be explained by the model, which is desirable. At 0.1139, although by this metric indicates that most of out data can be explained by the model, the linear regression model is not a good fit for this data. The adjusted R2 is very similar to the simple R2, which is at 0.1014 is also a desirable value for most of our data can be explained by the model. At 9.123, our F-Statistic is very low which, like the t-test, indicates the model is not very significant. Having a very small p-value like in this model is desirable, since the smaller the value the more significant the model in this case we can see it is significant but no so much **??**.



```
Call:
lm(formula = confirmed_cases ~ Current.BCG.vaccination., data = clean_ds)

Residuals:
   Min     1Q Median     3Q    Max
-36130 -16145  -4337  -2423 272200

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                 36650       8346   4.391 3.85e-05 ***
Current.BCG.vaccination.   -31089      10293  -3.020  0.00351 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41730 on 71 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.1139,    Adjusted R-squared:  0.1014
F-statistic: 9.123 on 1 and 71 DF,  p-value: 0.003507
```

Fig. 13. Summary of Linear Model to predict Confirmed cases using BCG vaccination feature.

III. Results

A. Modeling and Evaluation

After having a clear understanding of the features, their relationships and characteristics, we proceeded to filter one more time the data, since its been said by epidemiologist that certain thresholds are needed to be crossed in order to understand the COVID-19 spread, since we are dealing with a exponential phenomena. The filtering consist in analyze only countries with more than 3,000 confirmed cases and the country belongs to Upper middle income or , since we can observe in the previous plots like in figure 16 that actually there exist some trend but there exist too much density (countries) in the below are of confirmed cases.

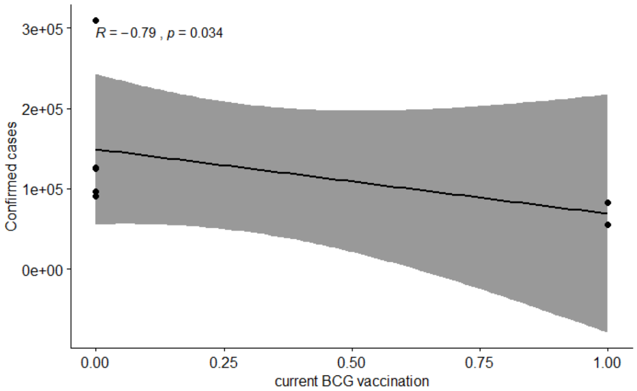


Fig. 14. Scatter plot and relationship by Spearman correlation Filter by 50k confirmed cases , High and Upper middle income .

The Spearman correlation test results imply that as it was observed before there exist some trend and some correlation betwwen the number of confirmed cases by country and its policies of BCG vaccination, for further exploration a lineplot in figure 15 is depicted in order to show the trends between countries with (in blue) and without (in red) universal policies of BCG vaccination .

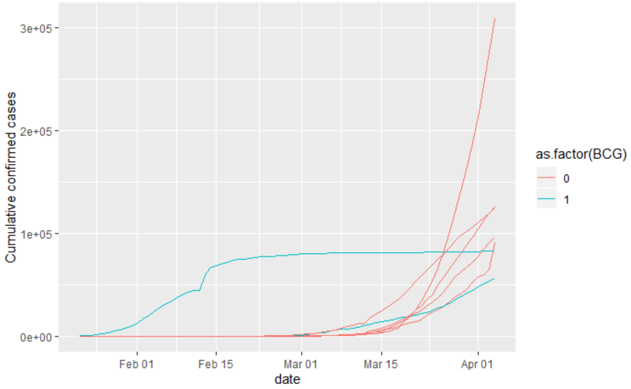


Fig. 15. Lineplot of confirmed cases daily per country .

IV. Discussion

As we can see in figure 16 a scatter plot between these two variables showing the values of Spearman correlation and denoting a negative relationship (-0.46); although it is clearly appreciable that the countries that currently have a vaccination policy tend to have fewer confirmed cases, there is a high density of observations with less than 50,000 cases, which causes the correlation to be biased. this indicates that it is possibly premature to measure the correlation in BCG vaccination vs confirmed cases around the world, or it is necessary to filter the cases to a number of confirmed cases greater than just 500, as this work proposes.

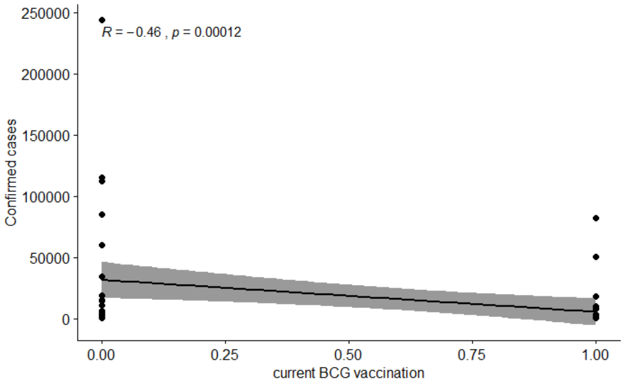


Fig. 16. Scatter plot between confirmed cases and BCG vaccination polices relationship by Spearman correlation.

V. Conclusion

Conclusion (What you learned from this exercise)

we have applied the CRISP-DM methodology from top to bottom, starting from a simple Exploratory Data Analysis (EDA), in order to understand the characteristics of the features and its relations among them, using methods such ANOVA, correlation test (Pearson and Spearman), modeling data for understanding using heatmaps to visualize groups and trends contained in the dataset, and fitted and evaluating the models using a simple linear model.

The results presented in this paper are very encouraging. First, it demonstrated for countries that have expressed observable trends (more than 50,000 confirmed cases) that BCG vaccination reduced the number of reported COVID-19 cases in a country, as shown in the spearman correlation test [**?**] this can lead to understand better this phenomena and eventually build a model that catches trends among countries.

References

[1] Scott Dylan. The covid-19 risks for different age groups, explained, 2020.

[2] María José Rubio Hurtado and Vanesa Berlanga Silvente. Cómo aplicar las pruebas paramétricas bivariadas t de student y anova en spss. caso práctico. *Reire*, 5(2):83–100, 2012.

[3] Koerth Maggie. How coronavirus tests actually work, 2020.

[4] Koerth Maggie, Bronner Laura, and Mithani Jasmine. Why it's so freaking hard to make a good covid-19 model, 2020.

[5] Aaron Miller, Mac Josh Reandelar, Kimberly Fasciglione, Violeta Roumenova, Yan Li, and Gonzalo H Otazu. Correlation between universal bcg vaccination policy and reduced morbidity and mortality for covid-19: an epidemiological study. *medRxiv*, 2020.

[6] Nidhi Prakash. Doctors and nurses say more people are dying of covid-19 in the us than we know, 2020.