**BUSINESS ANALYTICS**

Submitted by,

**Section – B | Group – 3 | PGPM 2020-21**

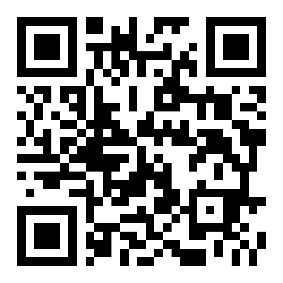
Abdullah Al Faruque (P202B046)

Dharun Babu K (P202B050)

Himanshi Saxena (P202B053)

Priya Dhiman (P202B059)

Sarthak Mahapatra (P202B062)



**Table of Contents**

(Hyperlinks are provided in each section for easy navigation)

**01** [**Introduction**](#INT) **2**

**02** [**Project flow**](#PF) **3**

**03** [**Approach used**](#AU) **3**

**04** [**Data cleaning, Data Preparation & Deliverables**](#DC) **4**

**05** [**Country wise data Interpretation**](#COUNTRY) **5**

**05.1** [**India**](#INDIA) **5**

**05.2** [**Sweden**](#SWEDEN) **9**

**05.3** [**New Zealand**](#NZ) **13**

**05.4** [**Australia**](#AUS) **18**

**05.5** [**USA – California**](#USA) **23**

**05.6** [**Italy**](#ITALY) **27**

**06** [**Overall Interpretation**](#OI) **31**

**07** [**Conclusion**](#CONC) **31**

**01** **Introduction**

“The Effectiveness of Lockdown on (controlling cases) COVID-19” – With a problem statement, we had analysed global data with respect to the objective to understand the actual impact of lockdown policies across various parts of the world in controlling the COVID cases. The lockdown was imposed on almost all countries from March onwards in a gradual manner depending on multiple factors like,

* Number of people
* People support
* Age group
* Economic conditions
* Individual Health index
* Country’s health index and facilities
* Availability of medical staffs
* People: medical facility ratio

For instance, countries like India are having low people: medical facility ratio, imposed lockdown from the very beginning stage of virus spread to reduce the rate of infection, and to build proper medical facilities to tackle the situation. With the provided dataset having parameters for lockdown stringency, economic policies, health policies, confirmed cases count, and confirmed deaths count, we can infer the relationships between different variables. Here is the gist of details from the dataset.

**COVID-19**

**Dataset**

**21\***

Key Indicators

**186**

Unique Countries

**01-Jan to 31-Aug**

Date range for analysis

**6** Countries considered for analysis

**3**

Newly added variables

**44**

Unique variables

**\*** - Excluding newly added variables

**02** **Project flow**

**Data collection**

**Approach finalization**

**Problem Understanding**

**Modelling**

**Data preparation**

**Data understanding**

**Interpretation**

**Deployment**

**Evaluation**

<https://github.com/Dharunbabu/Term2_BusinessAnalytics_Project> - Project details are available in the GitHub path

**03** **Approach used**

Considering the vast number of countries and complexities due to the absence of data for various countries, we approached this project by taking data of 6 different countries having positive or negative significance in either the number of cases or rate of infection or stringency indices to understand the lockdown strategies in controlling COVID cases. The countries which we took for analysis and modelling are,

* India
* Sweden
* USA (California)
* New Zealand
* Australia
* Italy

**04** **Data cleaning, Data Preparation & Deliverables**

Data cleaning and preparation are done to remove/modify irrelevant records to accommodate the model variables to get the results without ambiguity. Following are the steps are done to clean and prepare a dataset:

* Datasets are divided based on a country keeping only the countries in-scope for modelling
* Understand the relevant variables to be considered for the model. For ex., Economic policies are not needed. Irrelevant variables are removed from the dataset. Economic policy variables, Health policy variables (Except H1), Miscellaneous policy variable, Region Name, and code are removed.
* Added new columns like “Actual\_Date” for date conversion from text to date format, “NewCases” to calculate the newly confirmed cases added on that particular date, “RateOfInfection” to calculate the infection rate on any specific date comparing with the previous day
  + New cases = Confirmed cases on any specific date – Confirmed cases on previous date
  + Rate of Infection = New cases/ Confirmed cases on the previous date
* Split USA California data and UK England data from master data of USA and UK data, respectively
* Replace blank values with “1” in all flag variables of C1, C2, C3, C4, C5, C6, C7, H1 flags, only if they are having “0”. This is done because the “0” in the main variables C1, C2, C3, C4, C5, C6, C7, and H1 refer to no measures. So, if there are no measures taken, then it will be a generalized policy for the whole nation, so we had replaced blank with “1” in all flag variables.
  + “0” in main variables C1, C2, C3, C4, C5, C6, C7, H1 refers to “No measures taken”
  + “1” in main variables C1, C2, C3, C4, C5, C6, C7, H1 refers to “Generalised action” and “0” refers to “Targeted action”

**Deliverables**



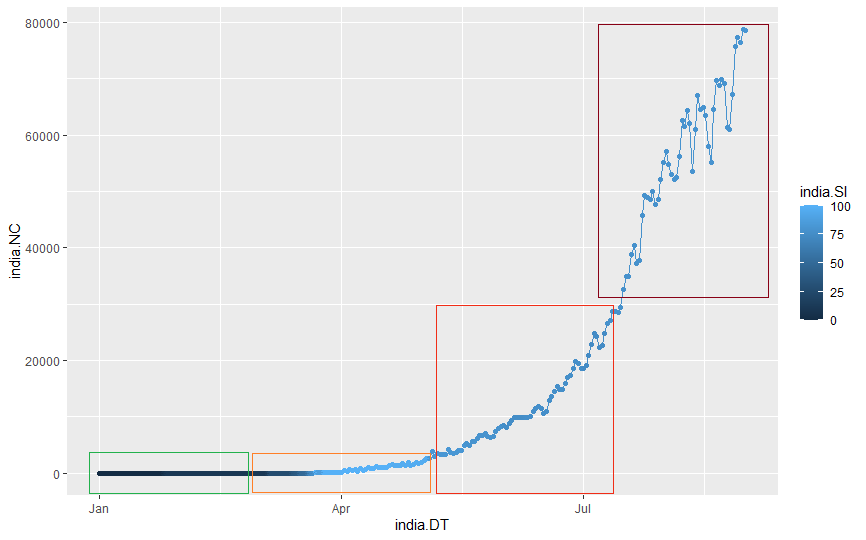
* ReadMe.xlsx – Files having details of all columns and policies details
* Section-B\_Group\_3.R – R-code of the project
* Section-B\_Group\_3.xlsx – Dataset considered for analysis
* Stringency Index – Calculation of Stringency index

**05** **Country wise data Interpretation**

**05.1 India**

India is one of the countries that was very proactive in implementing lockdown in the country. The nation went into lockdown from March 25th when the number of confirmed cases in the country was 562. In a country of 1.35 billion people, the country imposing strict lockdown with just 562 cases was a massive discission and was very different from what the western countries did.

If we look at some of the facts related to this, we find some interesting interpretations:

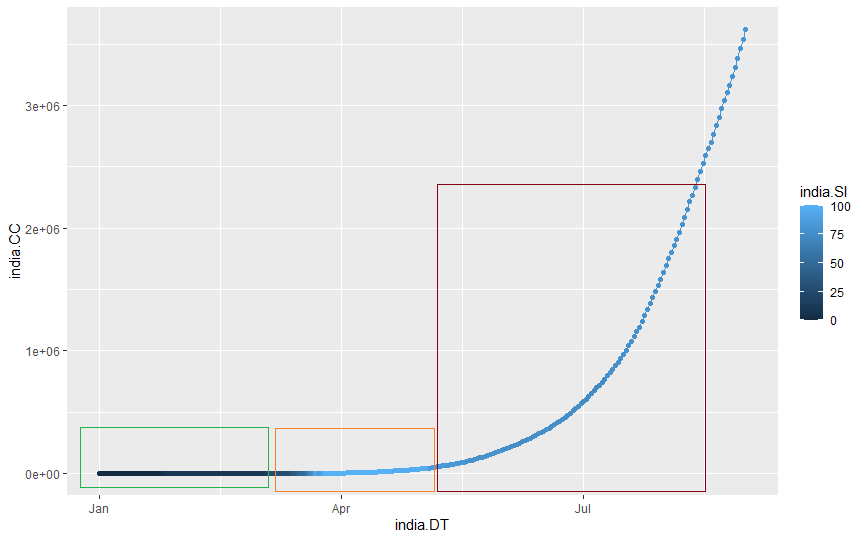


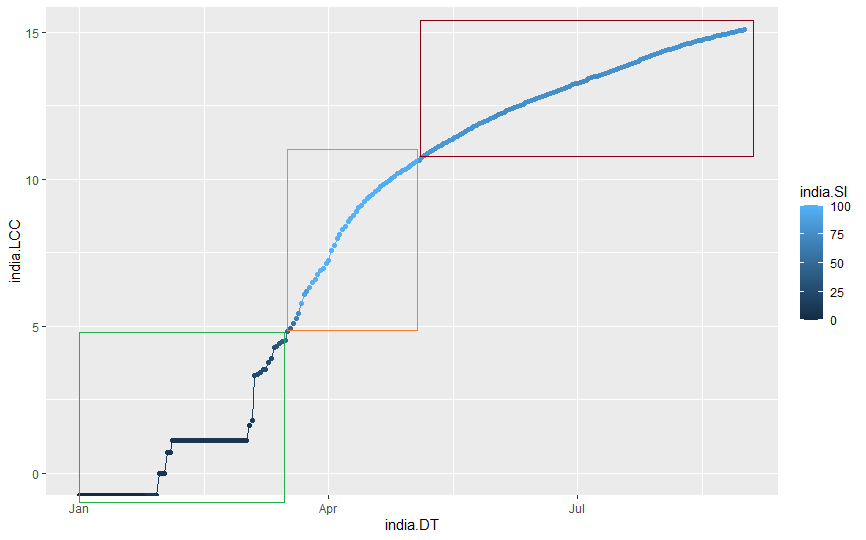
We find that the most stringent lockdown was imposed during April and May, during which we see that the number of new cases does not increase exponentially and was under control.

This clearly shows that if the lockdown is imposed early on, the infection rate can be kept under control. Post 31st May, India had to relive some restrictions to keep the economy going. The reduced stringency in the lockdown can be seen as associated with increasing new cases every day. As the lockdown was slowly reduced in stringency, the new cases started increasing at an alarming rate.

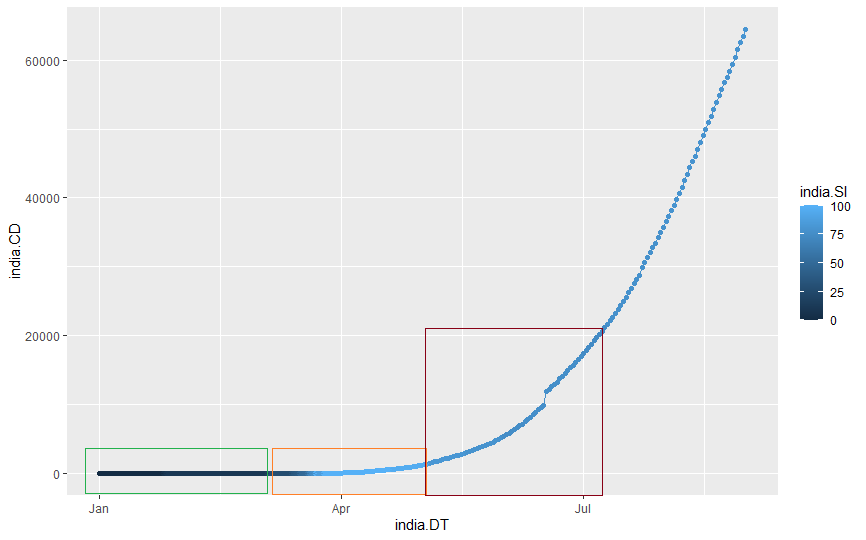
If we check the confirmed cases over time, along with the log of confirmed cases, we find that the infection spread at a much slower pace in India. The lockdown being proactive slowed down the pandemic and gave India time to prepare for the pandemic at its full force.

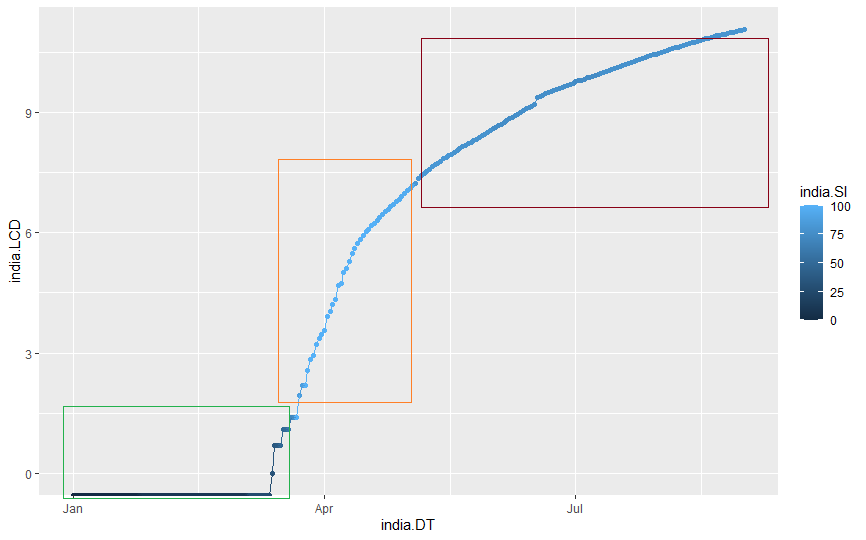
This also means that for a country of 1.35 billion people, the flattening of the curve of the log of confirmed cases was going to be delayed as well. We see that the graph shows that we have still not reached the peak of the pandemic even by August 31st.





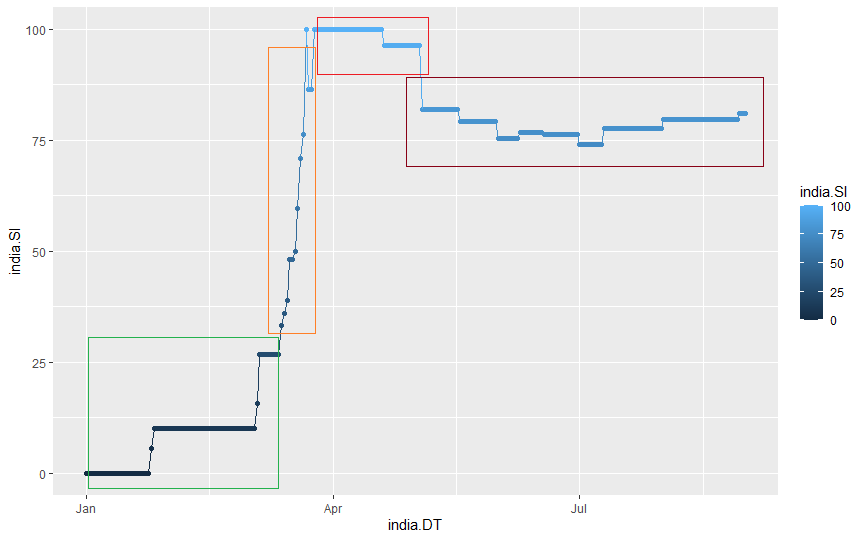
The graph comparing the deaths and log of confirmed deaths compared is also very similar to the confirmed cases graph. The early lockdown made sure that the people in India did not die in high numbers every day. The early lockdown allowed us to save millions of lives.





Stringency Index

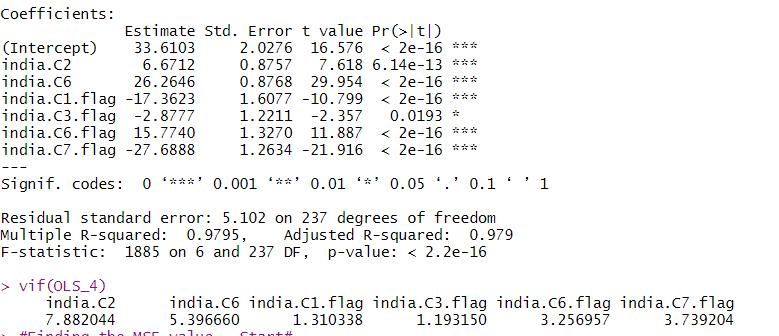
We find that the stringency index remains very high during the lockdown as the lockdown was very strict in its implementation in India. The lockdown was most strict during April and May. But the stringency index didn’t actually decrease much as India continues to fight COVID-19.



Now we have found all the factors taken into consideration to get the Stringency Index, but for India and its large population, we try to figure out the correct model using regression, removing the factors that do not play a major factor or have high multicollinearity.

* We first remove factors that have absolute multicollinearity in our first iteration.
* Next, we remove the factors that have high multicollinearity,
* And finally, we remove the factors that have a low impact on the model for India.

The final model looks something like this & gives us an idea of how stringent the lockdown in India has been, taking into consideration the most important factors.

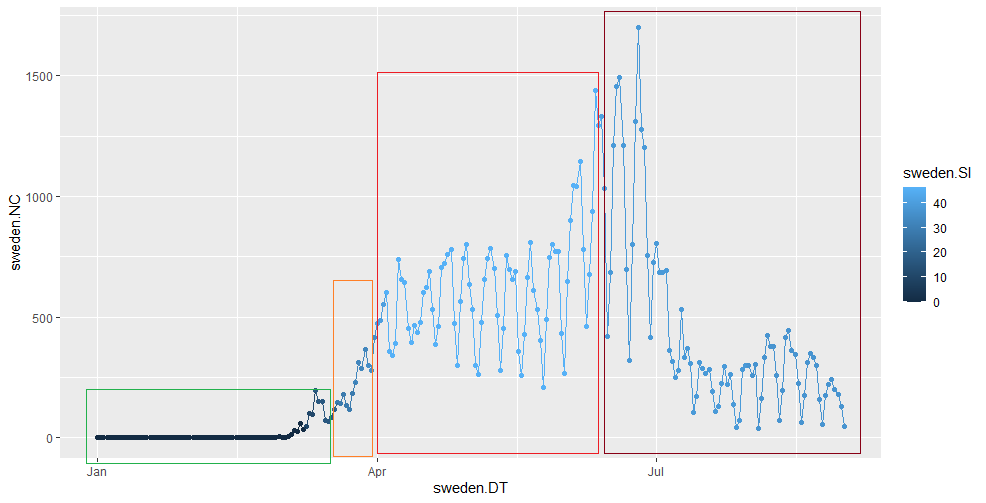


**05.2 Sweden**

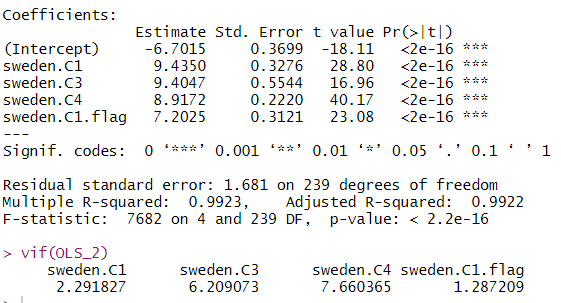
To tackle the pandemic, Sweden took an unusual strategy of not imposing any strict lockdown. Now the question is whether it was a good step or a failure. The data from the World bank is taken from the month of January 2020 till August 2020 to analyze the situation of the outbreak in Sweden.

The first step to proceed with our analysis is to clean the data. We filtered out all the nonessential figures from the whole Data, which are not required for the analysis like NA values. After we had clean data, we calculated the rate of infection and the Stringency Index for each day from the cleaned data.

The next step was to create the R code to analyze the data with a better aspect. The stringency index was taken as the dependent variable and other fields as the independent variable. On analyzing the graph for new cases date-wise, a zig-zag trend is seen. It can be said that the number of cases is increasing on alternate days and it reached the peak in the end of month June.



On the first linear run on train data with all the variables, it was found that half the variables were insignificant. After removing all the insignificant variables through multiple linear runs, we concluded with the following variables as the significant ones:



Sweden C1: School closing;

Sweden C2: Workplace closing;

Sweden C1. Flag:

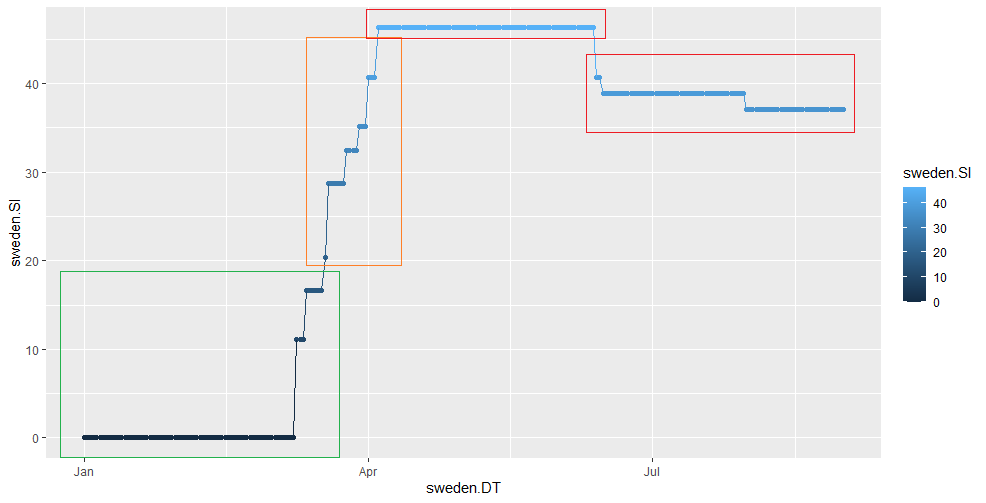
School Closing flag;

Sweden C4.flag:

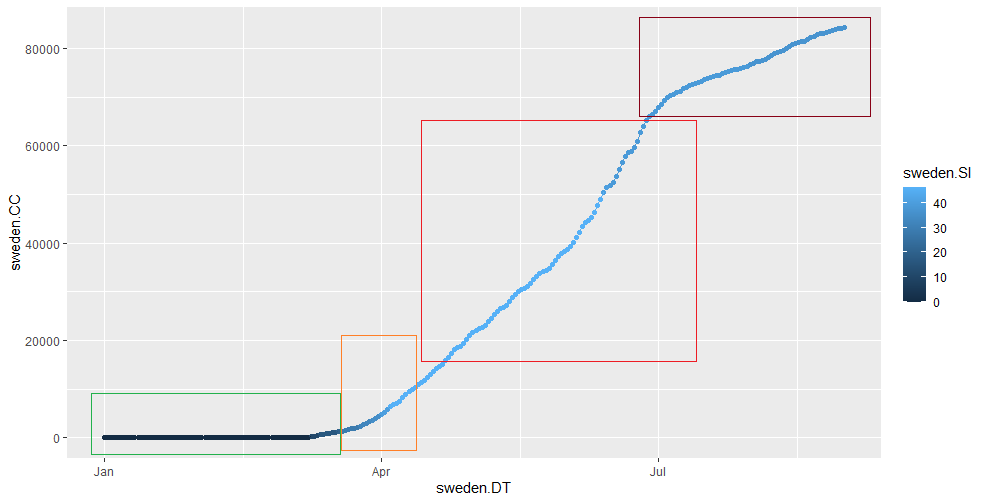
Restrictions on gathering flag

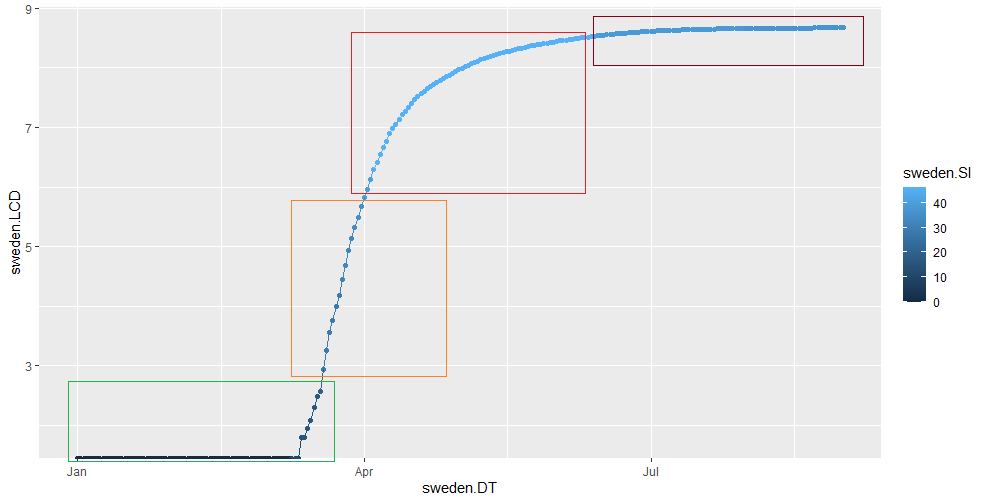
The mean square error was calculated as 612.7143 for these observations.

As per the Stringency Index graph, the actual lockdown started in the second week of month March and it was less than 50% which means no strict measures were taken to impose complete lockdown. Now, let’s look at the how the lockdown impacted the confirmed cases and death rate in the country.

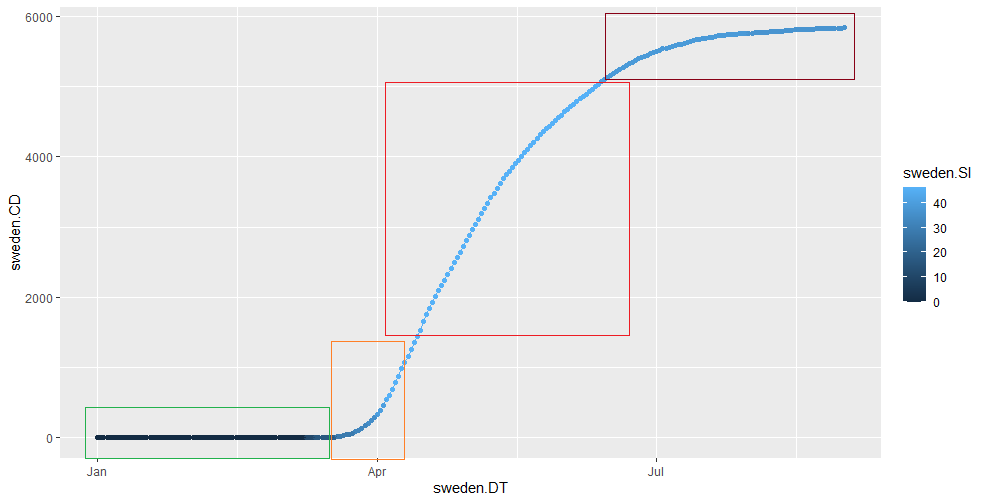


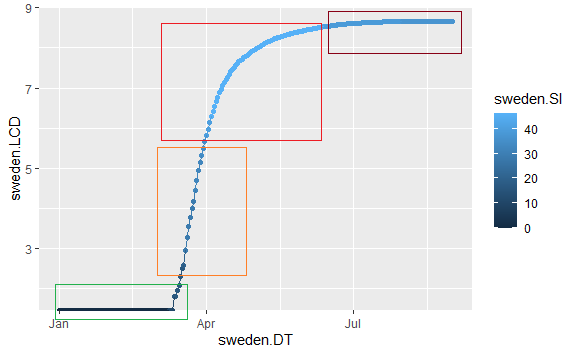
From the below graph of log of confirmed cases it was found that the number of confirmed cases increases at a very fast pace during the month of March to April and the strictness of lockdown also increased to control the cases. Later on, it was seen that the number of cases were increasing but at a very slow rate.





Sweden is an European country with high number of old aged people. Since corona severely affects the old aged, Sweden saw a high increase in the death rate during the initial days of Covid outburst. As the symptoms got clear and measures were taken to stop the impact of corona virus through social distancing, the rate of increase in the death rate was seen to be minimal after the month of July





**05.3 New Zealand**

The first case of COVID-19 was reported on February 28th 2020 in New Zealand. As of October 31st 2020, the country has had 1,957 cases (1,601 confirmed and 356 probable), of which 75 are currently active, and 25 people have died from the virus, with cases recorded in all twenty-district health board (DHB) areas. The pandemic peaked in early April, with 89 new cases recorded per day and 929 active cases.

*Important Dates*

1. National level high alert lockdown was introduced on **March 21st 2020.**
2. **From March 21st to May 14th** the lockdown restricted internal movements internally between cities and regions
3. **From June 8th**, the lockdown was completely removed internally while foreign travel remained restricted. Public transport, private gatherings, public events, schools, and universities were opened.
4. **On August 12th**, due to the possibility of a second wave, some restrictions were re-imposed (partial lockdown). Restrictions on gatherings were reinstated.

***Iteration with all variables***

*>* *Linear\_1 <-nz.master$StringencyIndex~nz.C1+nz.C2+nz.C3+nz.C4+nz.C5+nz.C6+nz.C7+nz.C8+nz.H1+nz.C1.flag+nz.C2.flag+nz.C3.flag+nz.C4.flag+nz.C5.flag+nz.C6.flag+nz.C7.flag+nz.H1.flag*  
  
During the first iteration to check the fit of the model, 4 variables returned were clearly found to be insignificant as they returned “NA” for the intercept values. These were discarded in the next iteration.

***Iteration to remove multicollinearity owing to high VIF value***

*> Linear\_2 <- nz.master$StringencyIndex~nz.C1+nz.C2+nz.C4+nz.C5+nz.C6+nz.C8+nz.H1+nz.C1.flag+nz.C2.flag+nz.C5.flag+nz.H1.flag*

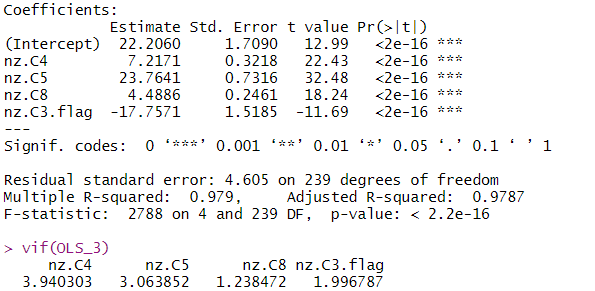
The second iteration returned a lot of variables with high VIFs. Variables with high VIFs indicate multicollinearity, and the variables of interest do not have high VIFs. Hence the iteration was run multiple times to remove multicollinearity from the model in this iteration.

***Iteration to arrive at final model***

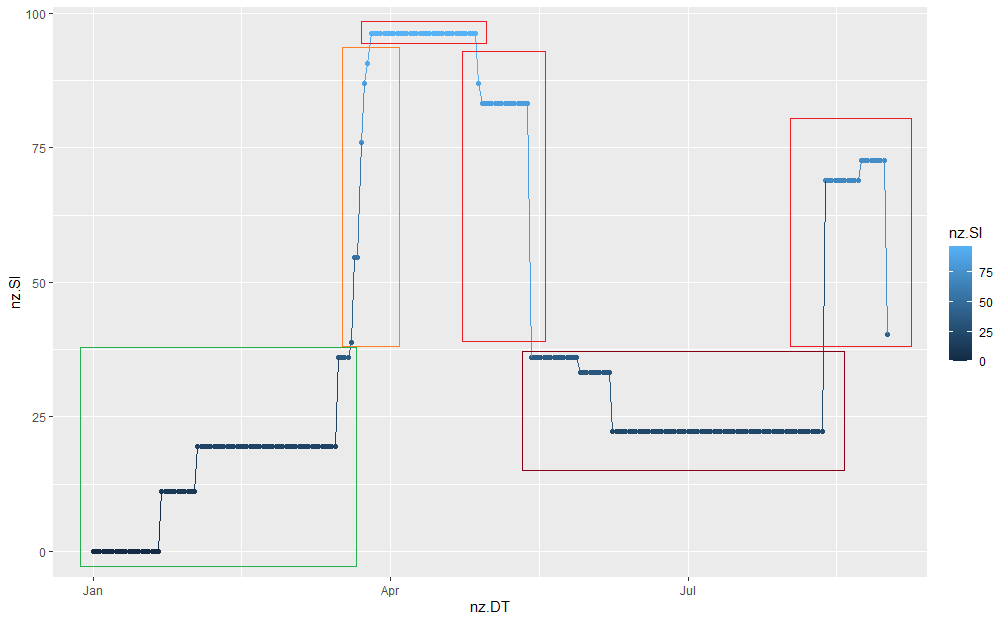
*> Linear\_3 <- nz.master$StringencyIndex~nz.C4+nz.C5+nz.C8+nz.C3.flag*

The third iteration returned our model analysis from the linear regression and all variables of interest were accumulated.

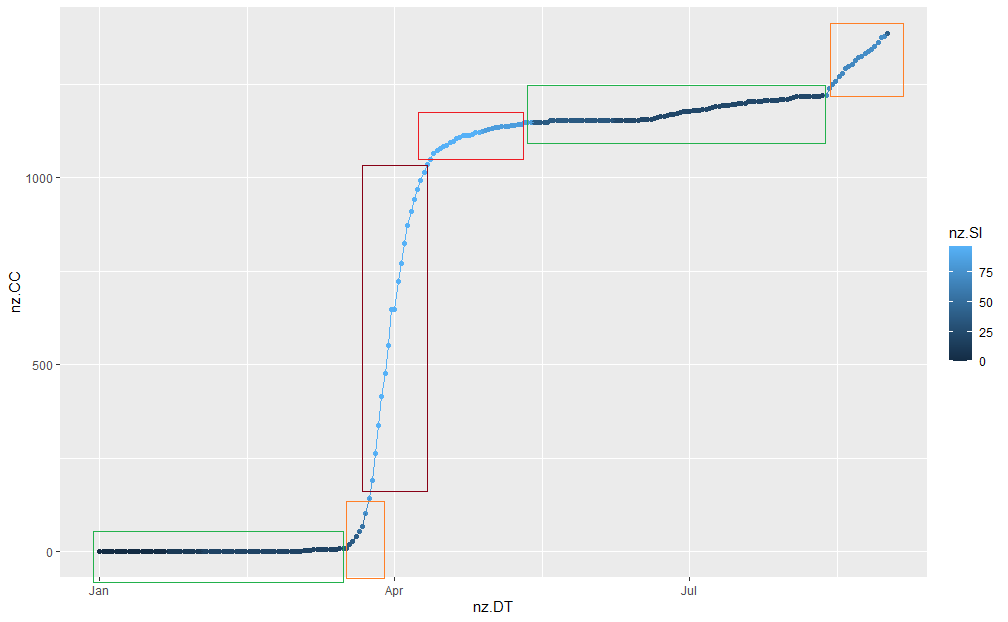
The model from the R code is as under: -

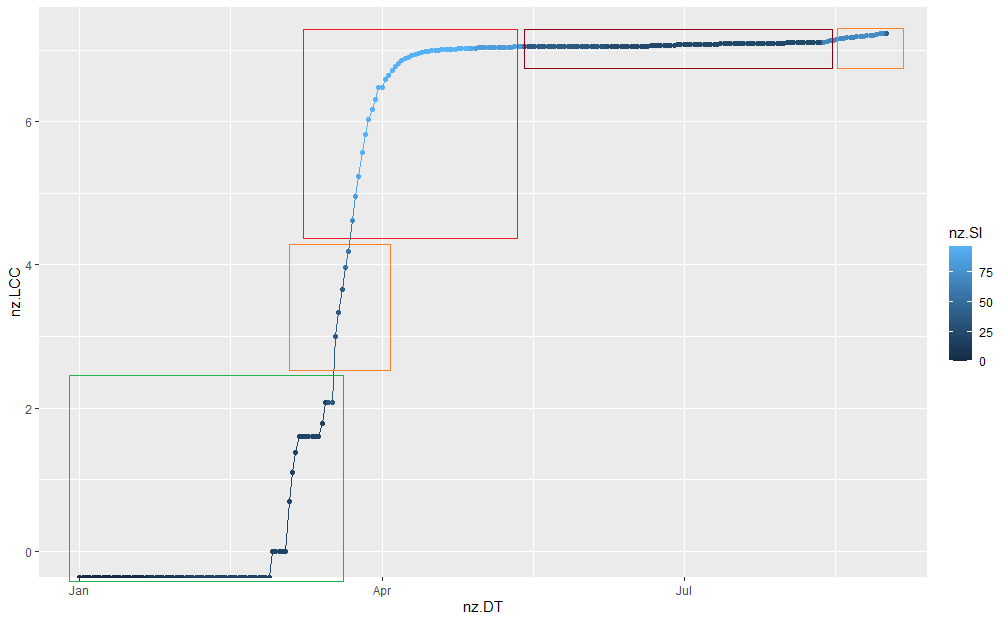


Graphical analysis of the final model:

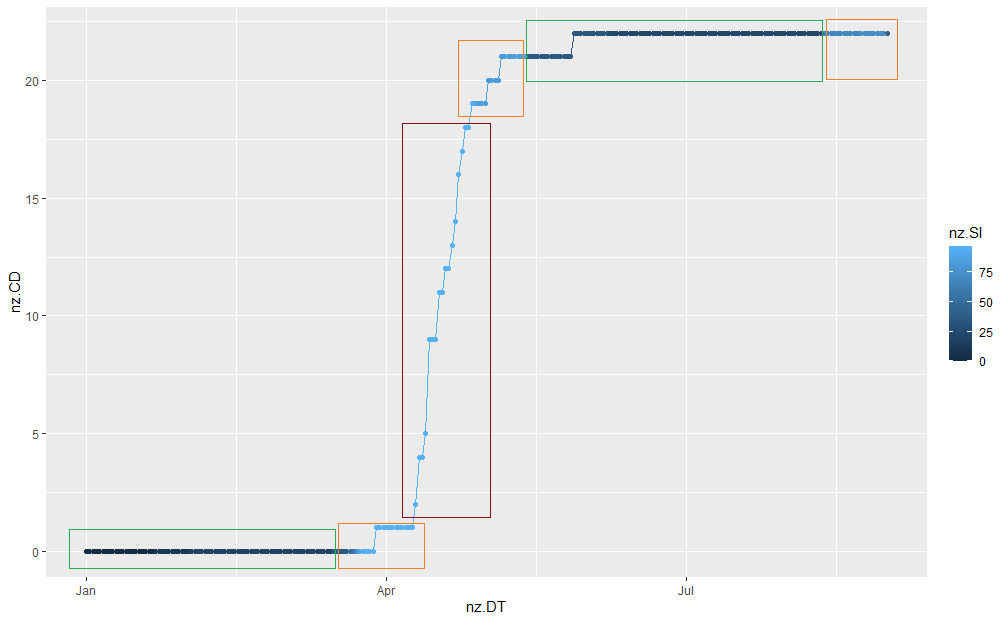


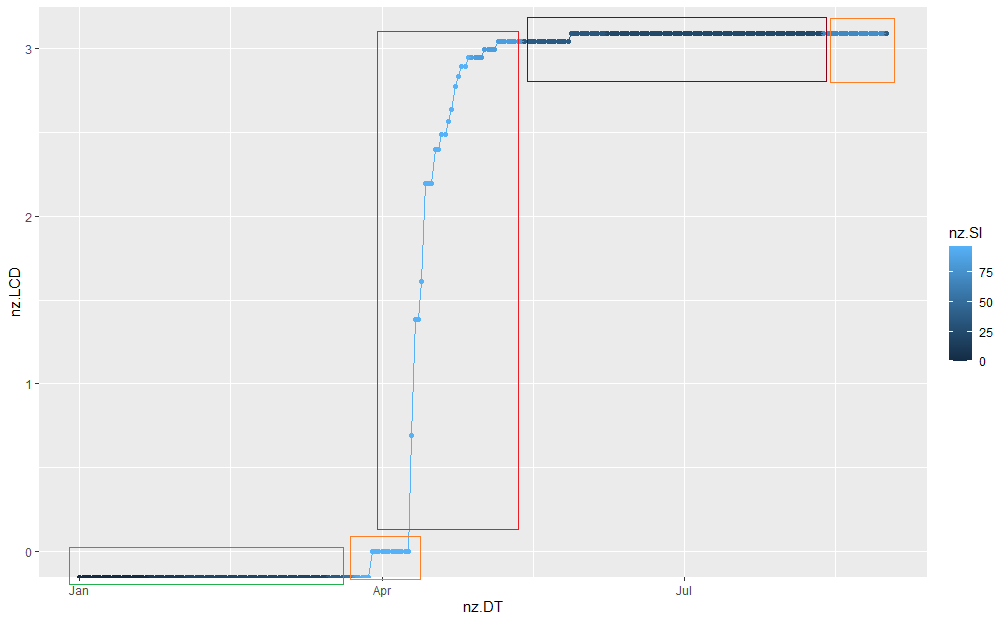
From the R-plot, it is evident that the stringency index was highest at the onset of the first lockdown. The strict lockdown phase continued till mid-May, and then tapering off of the stringency index is observed. The stringency index rises by a notable margin at the onset of the second lockdown, followed by a period of subsequent relief.





The confirmed cases shot up post lockdown as people who had contracted the virus pre-lockdown were now diagnosed at a higher rate. The exponential increase in confirmed cases was persistent till early May. The logarithmic plot of confirmed cases shows how the curve started flattening from May. The subsequent removal of lockdown and onset of the second lockdown has not affected the logarithmic scale too much, indicating an adequate level of control in the rise of COVID cases in New Zealand post May 14th. On the normalized CC vs. DT plot, we see a tiny spike at the onset of the second lockdown date.





The number of deaths saw an exponential rise during the lockdown period, and the curve started to flatten mid-May after the lockdown was removed. Both the graphs (logarithmic included) show that post-June, the number of deaths in New Zealand due to COVID has reduced and remains consistently low to date.

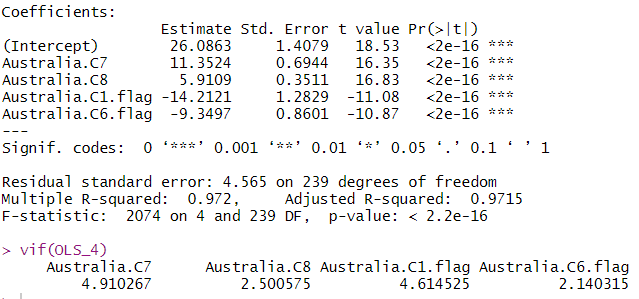
**05.4 Australia**

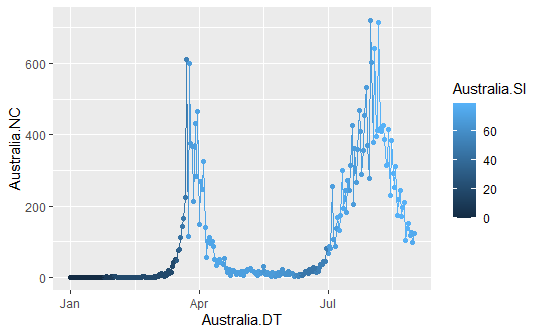
We have checked the linear relationship for the model. Stringency index is the indicator of lockdown in the countries. It is inversely proportional to the rate of infections in the countries, for the high value of stringency index the rate of infection is high when we consider variables of all the countries together for analyses. The value to stringency index may vary from country to country as per the measures implemented.

In case of Australia they implemented the measures in mid-march by imposing lockdown and since then the rate of infection increase to a certain extent and start decreasing from there, the details have been explained moving forward.

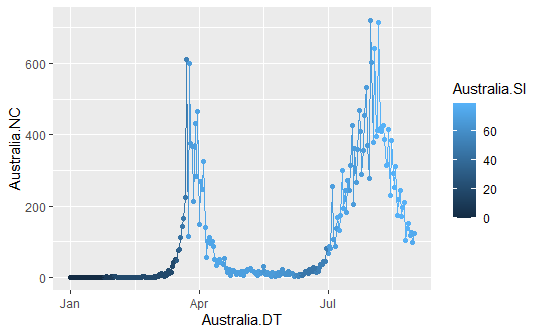
We have checked the fit of the model considering all the variables in 1st iteration which is ‘Linear\_1’ after analysing we removed the variables showing ‘NA’ in the 2nd iteration ‘Linear\_2’. Next, we checked for variance inflation factor values to check multicollinearity and consider removing the variable with highest VIF value as VIF closer to 1 will give a stronger model in iteration-3 ‘Linear\_3’. Then we checked and removed the insignificant variables from the model to get the best fit model in iteration-4 ‘Linear\_4’. The final variables used for the analysis in best model is:

Linear\_4 <- Australia.master$StringencyIndex~Australia.C7+Australia.C8+Australia.C1.flag+Australia.C6.flag

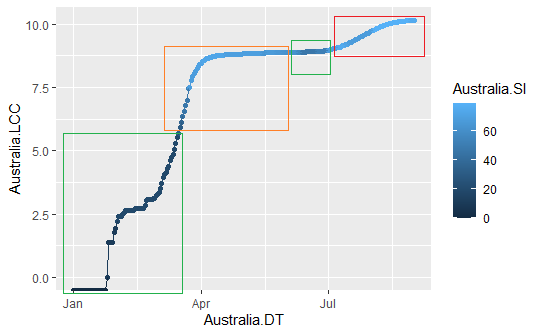




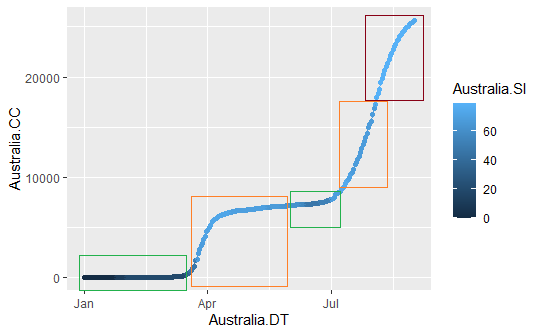
The above graph shows emergence of the new cases in Australia. The first confirmed case was identified on January 25th 2020 in Australia and the lockdown was imposed on March 20th 2020. The number of new cases started increasing rapidly after that around 350 cases per day was reported around March 22nd as shown in the graph the new cases are increasing from mid to end of March and started falling in the start of April to about 20 cases per day at the end of the month. The cases again started rising in June and ongoing till now.



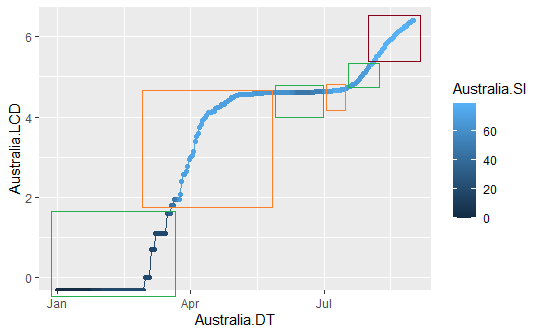
The above graph is plotted to show the effect of lockdown with respect to stringency index on new cases. We can see in the graph that the stringency index value is increasing in the March soon after the lock down imposed and number of cases started increasing and reaches to the peak which is continued till now.



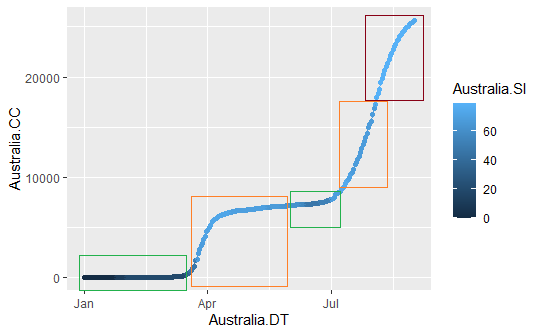
The above graph is plotted between the date and taking logarithm of confirmed cases. The graph shows the flattening of the curve from April showing the effect of lockdown (stringency index). The lockdown strategy in Australia has been successful as it helped to control the spread and emergence of new cases to some extent after the lockdown lowering down the count of new cases to 20 cases in the start of April as the graph showed flattening of the curve from the start of April.



The graph is plotted to show the effect of lock down on the confirmed cases. We can see that the confirmed cases are increasing from mid to end of March from around March 20th till the start of April after which the graph is flattened showing the constant in marginal change in the number of cases per day. The lockdown was re imposed in around June 30th after which the cases tends to rise again and ongoing till now.



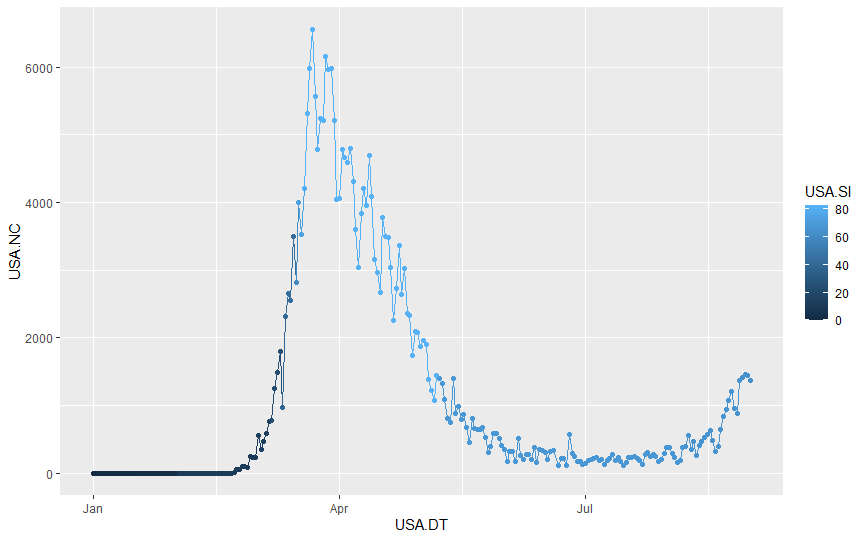
The above graph shows the flattening of curve for the number of confirmed deaths in the country after imposing the lockdown in Australia. The first death was reported on March 1st following the 2nd death on March 4th increasing the death toll 100 by the end of May. The current death count has reached to 907 in October.



The above graph is plotted to show the confirmed deaths with respect to the stringency index showing the effect of lockdown in the country. The 1st confirmed death was on March 1st and the lockdown was imposed on around March 21st and since than the number of deaths is increased to 100 by the end of May and increasing the count again after July making the count increased to 907 currently in October. The lockdown slightly stabilized the number of deaths as shown in the graph after May to mid-August, during which the lockdown was slowly started lifted but the number of deaths didn’t tend to decrease forcing the government to re impose the lockdown on around June 30th and the death count increasing after that and reached to 907 in October.

**05.5 USA – California**

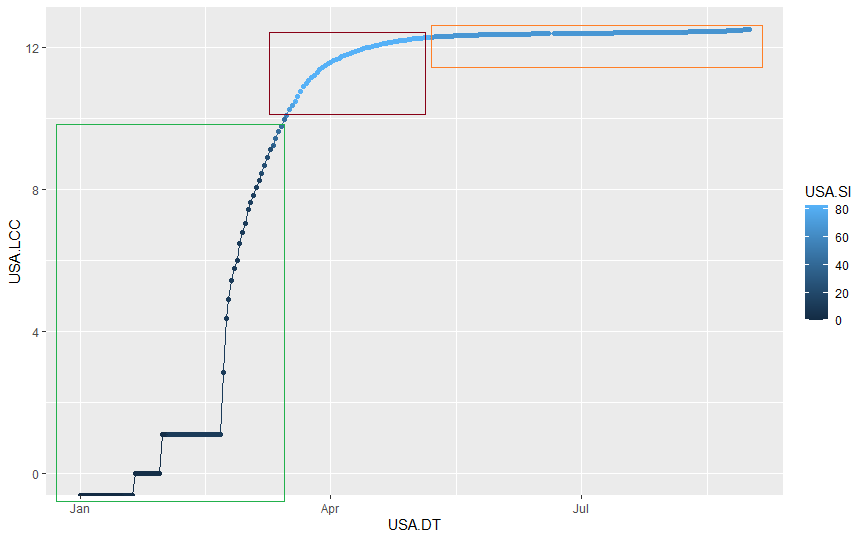
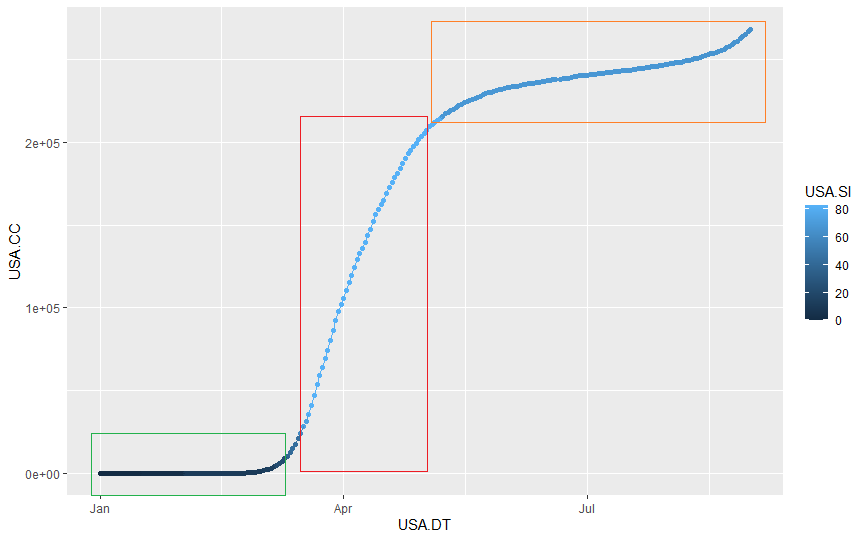
The United States had a varying degree of lockdown in different states at different times. So, for the purpose of this analysis we focus only on the data collected from the most densely populated state California. USA was very late in acknowledging the threat of the virus and very late in implementing the lockdown as well. The state of California implemented a full lockdown on March 19th by which time it was already having nearly 4000+ new cases per day and the number of cases per day were increasing by 1000 or more every day.



The delay in implementation of the lockdown made the numbers spiral out of control very quickly.

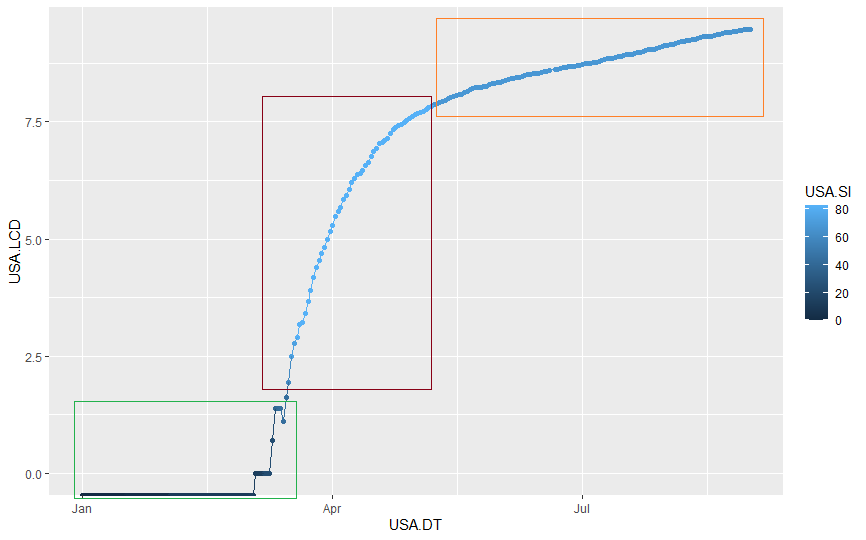
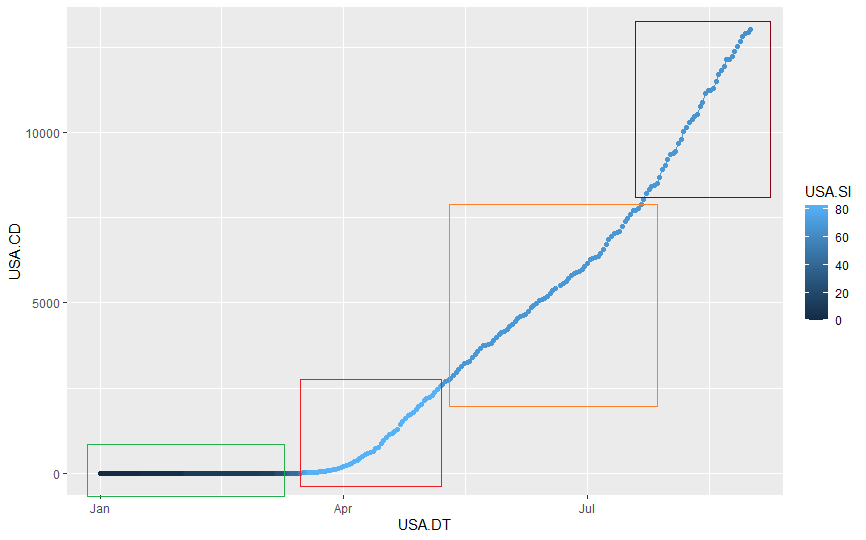
By 31st August the state on California only was showing 6000+ new cases per day and the count just kept increasing.

The lockdown stringency was reduced towards the end of May and June and the numbers again increased out of control. We saw the second wave hit again towards the beginning of July as the numbers started increasing again.



The curve that had somewhat flattened towards July started to rise slowly again as the number of new cases started to increase again.

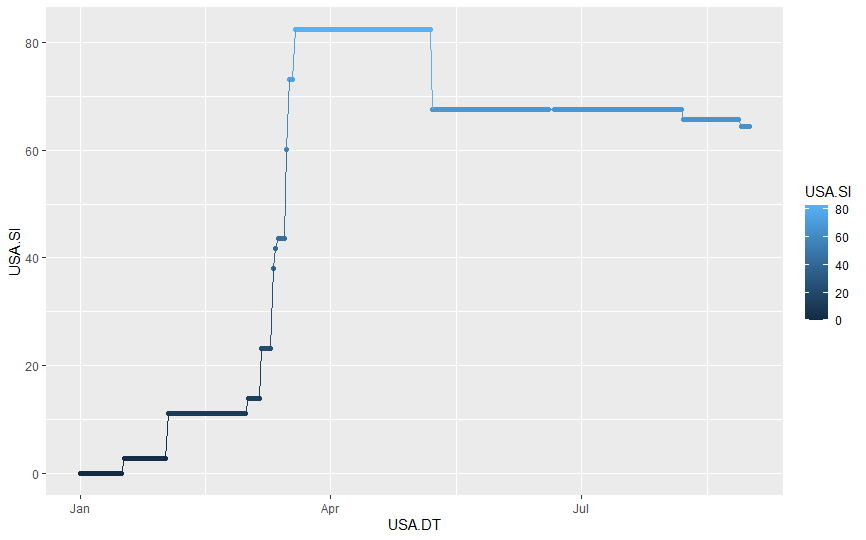
At this point the lockdown was more or less almost lifted with the country chaos due to the increasing number of cases.



The number of deaths graph showed a similar trend and with the numbers increasing rapidly at an alarming rate. There were over 10000 dead in California alone before the end of August as the count kept increasing with an increasing rate.

**Stringency Index:**

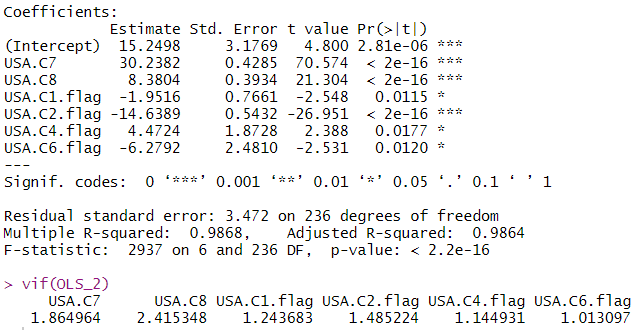
The stringency of the entire lockdown overall remained on the lower side except for a couple of months.



The delay in the implementation of the lockdown by approx. 2 weeks cost USA a lot of lives and uncontrolled number of cases. The stringency did decrease towards the end of MAY and the numbers started rising again. As we see whenever there is a slight decrease in the stringency index, we notice the subsequent days to have increased number of cases.

Now to figure out the main factors responsible for the Stringency index applicable to USA California we created a model based on various factors.

We removed all multicollinearity and we also removed the less important variable to find the below model to represent the Stringency Index of the lockdown in California.

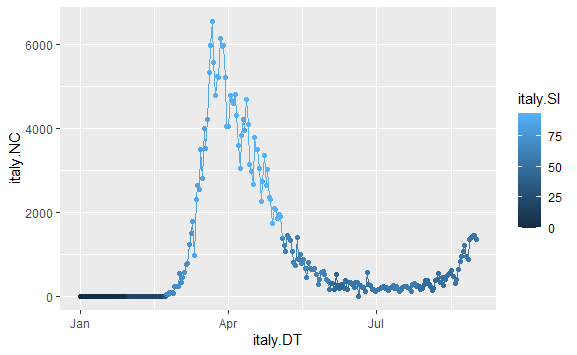


**05.6 Italy**

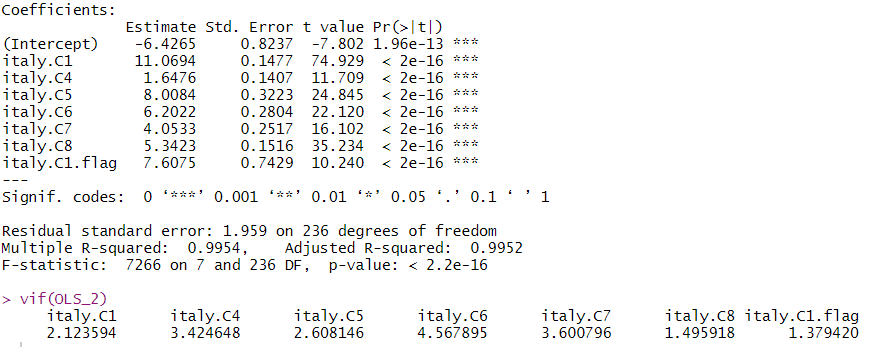
One of the top most impacted country in the world due to COVID pandemic is Italy. Being a highly admired tourist spot, a lot of people travelled from and to Italy which resulted in outburst of the virus in the early days of month Feb. New cases were found in the country as early as end of month Jan, but the government-imposed restrictions in the month of March. This delay has caused a lot of damage to the country. We further analyse the situation of Italy from the month of Jan 2020 till August 2020 through data received from World bank.

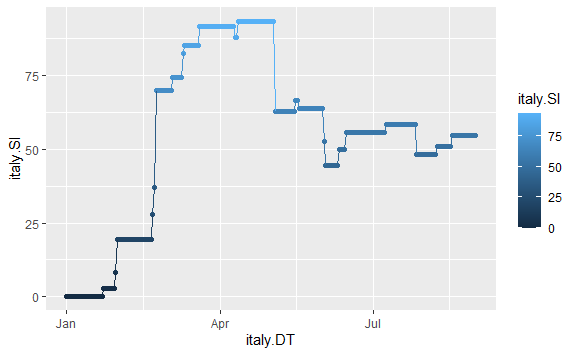
In our first step we cleaned the data and removed all the values like NA which were not required for the analysis. Once the data was cleaned, the rate of infection and Stringency Index was calculated for each day. To proceed further, we developed the R code specific to the country in order to analyze how the lockdown have impacted the number of cases. The Stringency Index was taken as the dependent variable and other fields as the independent variable.

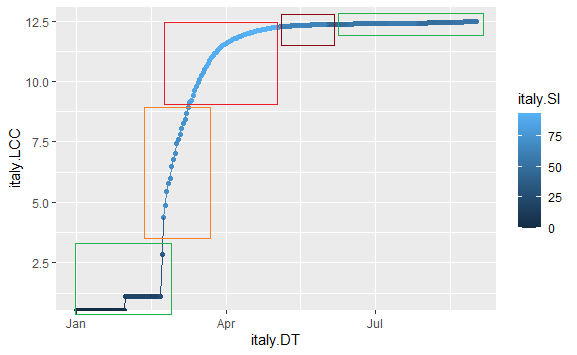
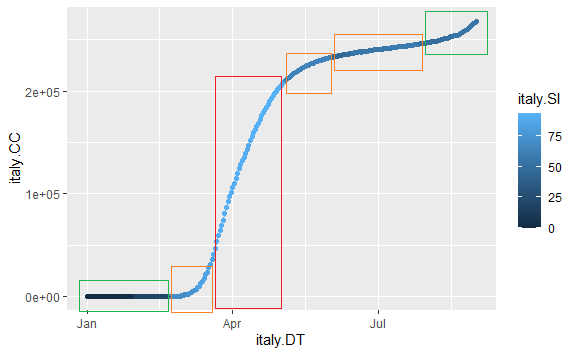
Through the graph below, it can be analyzed that the number of new cases the country reached its peak in the March month end. A fluctuation in number of new cases was seen in the April and May month but at a decreasing rate. But, a second wave of increase in number of cases can be seen in the month end of August.



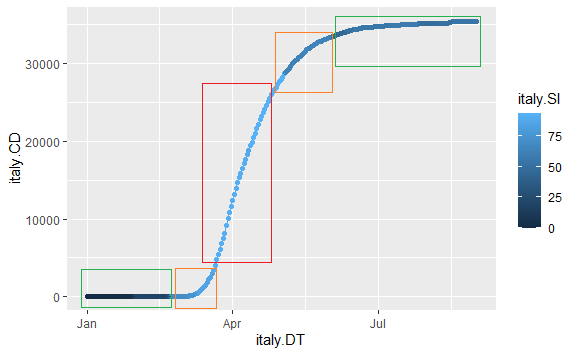
First linear run on the train data was done with all the variables, and it was found that only a few variables were insignificant. Once we removed all the insignificant variables through multiple linear runs, we concluded with the following variables as the significant ones:

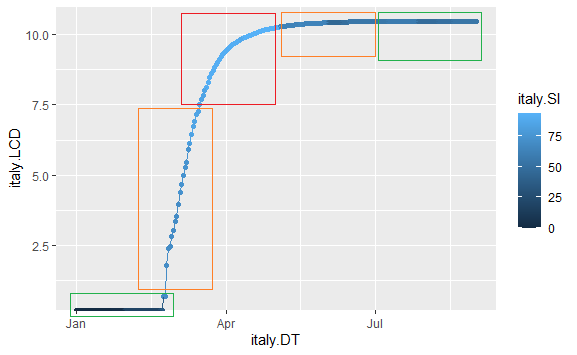


On analyzing the Stringency Index graph, it was found that the government took precautions early in the month of Feb. As the number of cases increased during the Feb and March month, the strictness increases to more than 75%. A slight relaxation was seen in the lockdown in the month of May, June and the lockdown measures were still continued till August month.****

From the below graph of log of confirmed cases it was found that the number of confirmed cases increases at a very fast pace from March to May end and the strictness of lockdown also increased during the period to control the cases. Post May there seems to be the stability in the growth of the confirmed cases and the strictness of the lockdown as we can see the flatting of the curve.****

Italy experienced a serious death rate starting from mid-March as shown in the below graph till May establishing the actual death numbers as 45,000 to 50,000. As the symptoms got clear and measures were taken to stop the impact of corona virus through social distancing, the rate of increase in the death rate was seen to be minimal after the month of July as the curve is slightly flattened after July. The strictness of lockdown was also increased from March to control the increasing death toll.

****

****

**06 Overall Interpretation**

From the models, it is very evident that stringency index is very related to the number of confirmed cases.

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | **Average Stringency Index** | **Average Rate of Infection** | **Remarks** |
| India | 54.71 | 12.25% | Proactive |
| Sweden | 28.60 | 6.26% | Stable |
| New Zealand | 39.83 | 4.44% | Proactive |
| Australia | 50.18 | 5.57% | Proactive |
| USA - California | 52.62 | 8.4% | Reactive |
| Italy | 54.71 | 12.25% | Reactive |

* Proactive Strategy – Countries reacting very soon by imposing stringent lockdown policies to control the rate of infection
* Stable Strategy – Countries believing in herd immunity imposed partial lockdowns by closing the places that might give rise direct infection Ex. Public transport
* Reactive Strategy – Countries imposing severe lockdowns after the confirmed cases were spiked up.

**07** **Conclusion**

With the help of regression models, we found stringency index is collective parameter of all critical variables of lockdown measures in one way or other. From the results, we can see that the rate of infection is directly/inversely proportional to rate of infection depends on the stage at which lockdown is imposed. If the lockdown is imposed after the cases spiked up, then by that time lots of clusters were created to further bubble up the infection. But if the lockdown is imposed in earlier stages, the number of direct infections is reduced to a greater extent which further reduce the rate of spread. Also, if any country is lifting its lockdown or leveraging the policies, almost all countries are again seeing the number going up rapidly.

Thereby, we are concluding the lockdown and the stage at which it is imposed collectively have impact over the rate of infection.