Case Studies in Data Science (COSC2669)

Final Report
WIL Project 19 (Data Buzz)

Group 19

Nitin Tundwal, Christangel Fargose, Mark Pereira, Shonil Dabreo, Hemanth Kumar

RMIT University

Nitin (s3572060@student.rmit.edu.au)

Christangel (s3794800@student.rmit.edu.au)

Mark (s3797413@student.rmit.edu.au)

Shonil (s3835204@student.rmit.edu.au)

Hemanth (s3823997@student.rmit.edu.au)

21 October 2020

Tables of Contents

Abstract

The corona virus has had a devastating effect on the global economy and health of the people affecting the lives of billions of people and almost all countries around the world. This pandemic has affected many businesses around the world leading to unemployment, poverty and shifting of business locations. This change has significantly impacted the economies and GDP of all the countries leading to global economic crisis. Businesses are trying to move their units in the countries whose GDP is not affected that heavily or in the countries where the recovery rate of corona virus is high. To assist the business users, make this decision we have developed a dashboard that takes the dataset as input and provide insights and predictions regarding corona virus recovery rate and effect of GDP on that country. We have developed a model which takes the dataset as input and based on the filters it will provide the insights to the user which will help him take important business decisions. Further enhancements include providing more significant insights and including more attributes which can improve the accuracy of the model's prediction.

1 Introduction

1.1 Purpose

The primary purpose of this report is to describe the motivation, scope, approach and functionality of a data science-driven web application to measure the recovery rate of each country and analysing the impact of Covid-19 on the GDP of the countries.

The secondary purpose is to describe the collaborative approach to coordinating work as a team to develop the application.

1.2 Coronavirus and its impact on the world

The COVID-19 pandemic has led to a dramatic loss of human life worldwide and presents an unprecedented challenge to public health, food systems and the world of work. The economic and social disruption caused by the pandemic is devastating tens of millions of people are at risk of falling into extreme poverty, while the number of undernourished people, currently estimated at nearly 690 million, could increase by up to 132 million by the end of the year [2].

Millions of enterprises face an existential threat. Nearly half of the world's 3.3 billion global workforce are at risk of losing their livelihoods. Informal economy workers are particularly vulnerable because the majority lack social protection and access to quality health care and have lost access to productive assets. Without the means to earn an income during lockdowns, many are unable to feed themselves and their families. For most, no income means no food, or, at best, less food and less nutritious food [2].

Millions of workers – waged and self-employed – while feeding the world, regularly face high levels of working poverty, malnutrition and poor health, and suffer from a lack of safety and labour protection as well as other types of abuse. With low and irregular incomes and a lack of social support, many of them are spurred to continue working, often in unsafe conditions, thus exposing themselves and their families to additional risks.

Further, when experiencing income losses, they may resort to negative coping strategies, such as distress sale of assets, predatory loans or child labour. Migrant agricultural workers are particularly vulnerable, because they face risks in their transport, working and living conditions and struggle to access support measures put in place by governments [2].

As of today, around 41,015,139 people were infected by coronavirus out of which 1,128,722 people died and 30,612,894 people have recovered so far. Currently there are 9,273,523 active coronavirus cases out of which 9,196,702 are in mild condition whereas 76,821 people are in serious condition. The coronavirus has affected around 215 countries and territories around the world. At the moment many countries have taken measures to slow down the spread of coronavirus. Many countries have declared restrictive measures such as lockdown, stay at home to contain the pandemic at a local level.

1.3 Coronavirus and its impact on countries economy and GDP

The coronavirus has affected the global economy and the GDP of the countries in a severe way.Early estimates predicated that most major economies will lose at least 2.4 percent of the value their gross domestic product (GDP) over 2020, leading economists to already reduce their 2020 forecasts of global economic growth down from around 3 percent to 2.4 percent. Global attitudes about the state of the economy amid the coronavirus are more negative in some countries than they were during the great global recession [6].

In some countries the economic downturn has indeed been extremely severe. In Spain, UK and Tunisia, the output of the economy in the second quarter was more than 20% smaller than in the same period last year. This is 4 to 5 times larger than any other quarterly fall on record for these countries. And in Peru the year on year fall was even larger, at 30%. In other countries, however, the economic impact has been much more modest. In Taiwan, GDP in the second quarter of 2020 was less than 1% lower than in the same period in 2019. Finland, Lithuania and South Korea all saw falls in their GDP of around 5% or less [3].

Countries which suffered the most severe economic downturnslike Peru, Spain and UK are generally among the countries with the highest COVID-19 death rate.And the reverse is also truecountries where the economic impact has been modest like Taiwan, South Korea, and Lithuania have also managed to keep the death rate low [3].

Growth could be weaker if downside risks materialise. In the near-term, the major downside risk is that the impact of the coronavirus proves longer lasting and more intensive than assumed in the projections. In the event that outbreaks spread more widely in the Asia-Pacific region or the major advanced economies in the northern hemisphere, the adverse effects on global growth and trade will be much worse and more widespread. Illustrative simulations of this downside risk scenario suggest that global GDP could possibly be reduced by 1½ per cent in 2020, rather than by ½ per cent as in the base-case scenario [7].

Unless the pandemic is stopped, economies and markets around the world will continue their free fall. But even if the pandemic is more or less contained, overall growth still might not return by the end of 2020. After all, by then, another wave is very likely to start with new mutations; therapeutic interventions that many are counting on may turn out to be less effective than hoped. So, economies will contract again, and markets will crash again [7].

1.4 Problem Statement

Measuring the rate of recovery of each country and measuring the impact of COVID-19 to each country's GDP. As it is well known to each one of us, this pandemic is worst one in decades. Loss of lives and hope has become a norm. The businesses are forced to shut, people are struggling for basic human needs. An evening walk is a leisure now. This all has proven very bad for the businesses. Global economy is collapsing. To absorb this loss, people are now forced to work in these difficult times, risking their lives to survive.

Due to these circumstances, the businesses are shifting their units to other countries which are not worst hit by the pandemic. To make a smart analysed decision we have come up with a solution which will help the businesses to make smart and informed decision.

We aim to identify the recovery rates of the country which will in turn tell us how well the country is coping with the pandemic. If a country has a better recovery rate than other, it means that country has taken certain fruitful methods which lowered their death rate and in turn that country will be opening their market sooner than others.

The world has taken a colossal impact in Economical sector. The WHO estimates that global labour income has been pushed down by more than 11% causing more than 3.5 trillion us dollars in losses. This constitutes of more than 5.5% of the world total Gross Domestic product (GDP).Companies around the world have been fearful about their survival in the pandemic because of the closure of the border in many countries and also for the companies who want to invest and expand their business have been worrying about their investment since they have no clue on where and when it should be done.

We are aiming to get the measures of each country's quarterly GDP including first two quarters of 2020. This will give us an estimate of country's financial condition. Also, this will help in identifying the countries which will recover faster than the others.

So, these metrics will be provided to the business owners which will help to select the country in which they can advance their business without much loss.

2 Method

2.1 Choice of variables and datasets.

To measure the recovery rate of each country we have used the coronavirus dataset which contains the data of coronavirus cases of different countries. This dataset is sourced from < https://github.com/CSSEGISandData/COVID19/tree/master/csse_covid_19_data/csse_covid_19_time_series > and is maintained by John Hopkins University. This coronavirus is a live dataset and is updated on a daily basis. The coronavirus dataset consists of the following data:

- Date- Date for the daily number of cases.
- Country- The name of the country in which the cases are detected.
- Lat and Long- Latitude and Longitude coordinates of the country.
- Type- Different type of cases i.e. Confirmed, Death or Recovered.
- Cases- The total number of cases [8].

To analyse and calculate the effect of the coronavirus on the GDP of the country we have used the Countries GDP dataset.
This dataset is sourced from < https://stats.oecd.org/index.aspx?queryid=33940 > and is maintained by Organisation for Economic Co-operation and Development. The GDP dataset consist of the following information:

- Location: Abbreviation of the country.
- Country: The name of the country.
- Measure: Shows the economic growth rate compared to previous quarter or same quarter of the previous year. It is either GPSA or GYSA.
- Frequency: Represents whether the value is annual or quarterly.
- Time/Period: Represents the year and the quarter of that particular year.
- Value: The GDP value [8].

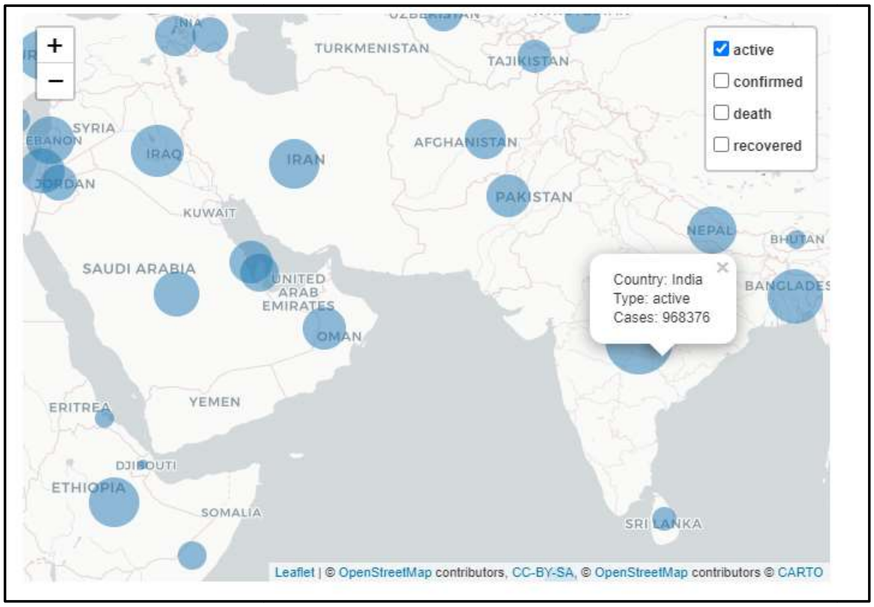
2.2 Data Preparation and Manipulation

Both the datasets were imported in the R using the base R function. The datasets contained missing values as well as attributes which were not required for the analysis. The data types of all the attributes were checked and converted if they were not in the required type. For the GDP dataset we have only included the data from 2016 because to show the effect of coronavirus on GDP the data before 2016 wouldn't make any significant difference [8].

The 'province' attribute in the coronavirus dataset was removed as it was not required in the analysis part. Similarly, the 'flag code' attribute was removed from the GDP dataset because it was not required for the analysis part. Both the coronavirus and GDP datasets were merged using left join based on a common attribute. We scanned for missing values in the merged dataset. There were negative values in the Recovered, Death, confirmed cases attributes which were removed. Confirmed cases have negative values which means no of patients tested and found are not confirmed so those values would be 0. Recovered cases have negative values which means no of patients tested again could be covid positive (confirmed).
Death cases have negative values which means it could be possible that negative deaths are patients who are not dead are still alive with covid severe condition(confirmed). We also checked for outliers in the dataset. The dataset is now clean and ready for the analysis and modelling part [8].

2.3 Analysis and Visualisations.

Our main aim is to help companies make decision regarding the country in which they can set-up their businesses during coronavirus. For that purpose, we need to analyse the coronavirus affected and recovered cases in that particular country and also the effect of coronavirus on the country’s GDP. To propose the solutions for decision making we have plotted the following visualisations:

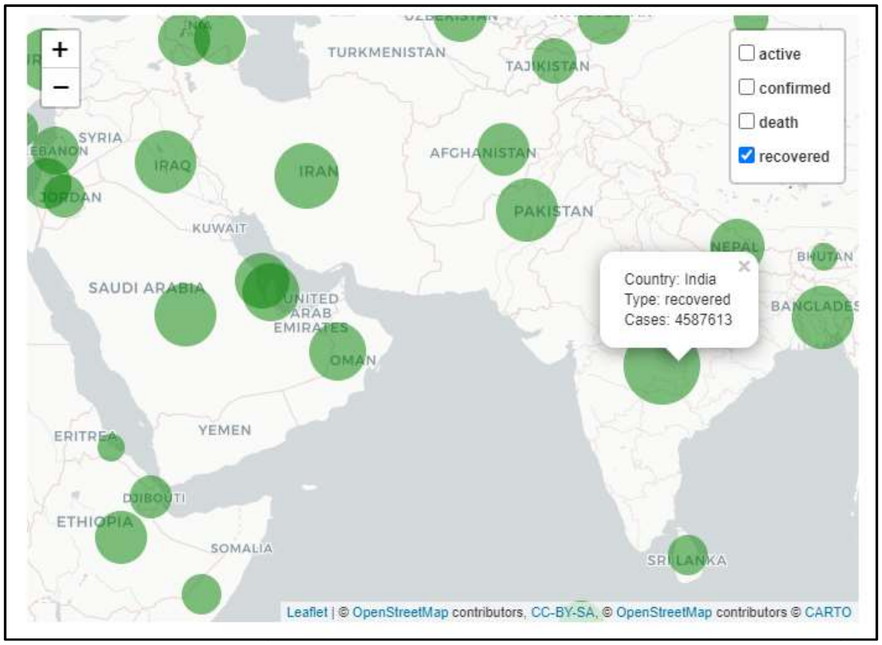


☑ active
☐ confirmed
☐ death
☐ recovered

Country: India
Type: active
Cases: 968376

Fig 1.1

We start by first plotting the interactive world map which shows different types of cases and their counts. We could know how big the numbers of cases are by the size of the circles shown for each country. We can also get value of cases count by navigating to country locations in map and according to the type selected, we get to know the no of cases for that type of a particular country. In the Fig 1, as we can see the no. of active cases in India are 968376.



☐ active
☐ confirmed
☐ death
☑ recovered

Country: India
Type: recovered
Cases: 4587613

Fig 1.2

Similarly, we can also get the no. of recovered cases in India which is 4587613. This means that the recovery rate is higher in India.

Through these visualizations, any business could know the total cases for each country since the beginning. This will give a glimpse of the countries that have managed to eradicate the covid. For e.g., if the numbers of active cases (circle size) are low and the recovered case are higher than all the other cases. The opposite of this could mean the countries that failed to control the covid.

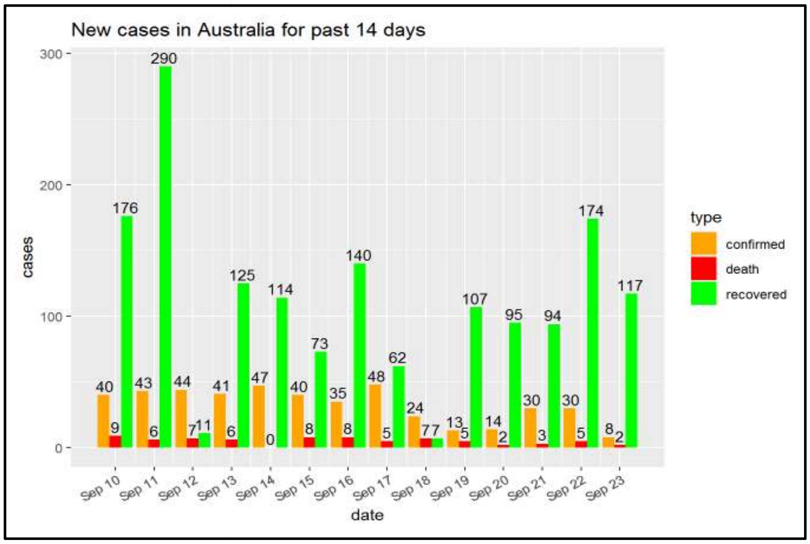


Fig 2

To understand how rapidly the cases are increasing or decreasing, we plot the daily new cases in Australia for past 14 days. As we can see in the Fig 2, overall, we could see that the daily recovered cases are higher than the confirmed and death cases. The no of recovered cases were higher among all the cases on 11th September and there were no deaths on 14th September. Considering the total cases, we found that the total no of cases were highest on 11th September and were lower on 18th September. Moreover, the death and recovered cases were similar on 18th September.

We also get to know that the increase/decrease rate of confirmed and recovery cases are almost similar but on different levels (Linear). Whereas, death rate is roughly stable. This means that no. of active cases would be stable and it suggests that the entrepreneurs should hold regarding their decision to open their business.

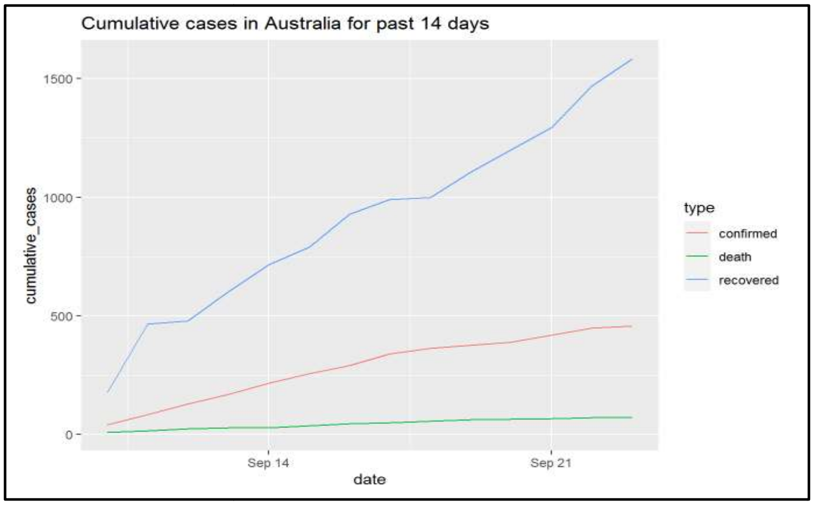


Fig 3

To understand how the cases have increased overall, we summed up the previous cases to today's cases (i.e. Cumulative) in Australia for past 14 days. As we discussed earlier, recovered cases have a linear relationship which means that the numbers are gradually increasing. Since there was less no. of deaths the line is flat. Whereas, confirmed cases tends to increase gradually till 17th September and then the total no. of cases have increased at a slower rate.

This type of graph illustrates how effective are the measures undertaken to control the spread of virus. Through the graph, we get to know that the control measures are slowly becoming effective. Thus, this suggests that the chances of Australia government releasing lockdown restrictions in coming weeks are higher and that the businesses should be ready to begin with.

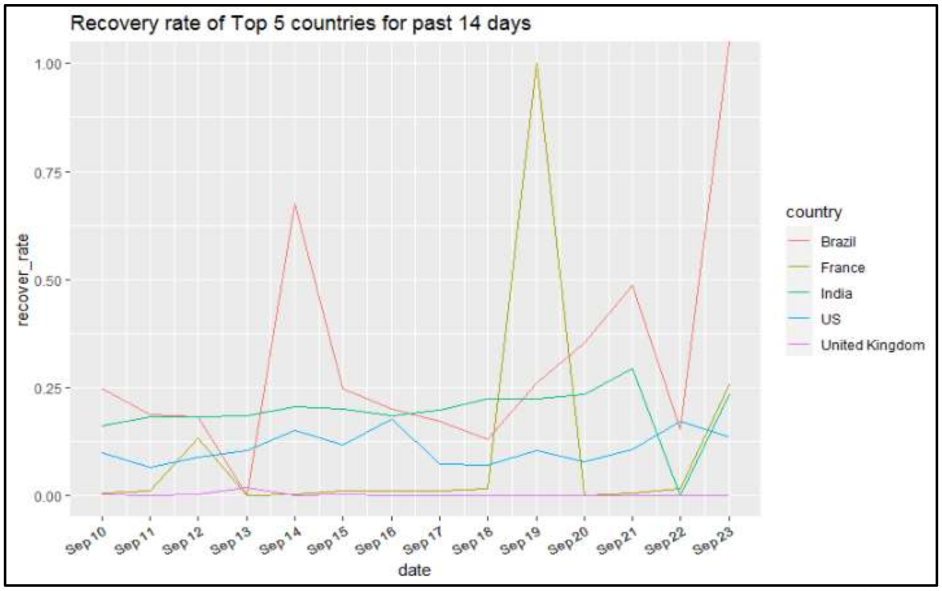


Fig 4

Finding the recovery rate in countries with lower confirmed cases would obviously be higher. So, as we can see in the Fig 4, we chose to find the recovery rate (in percentages) with respect to the daily confirmed cases of 5 countries with most number of cases for past 14 days. The recovery rate of UK is stable over time and is close to 0 percentile i.e. the no of confirmed and recovered cases daily are similar in UK. France and Brazil has relatively unstable recovery rate whereas US and India have around stable recovery rate throughout.

This means that though the no of confirmed cases in India are slowly increasing the recovery rate tends to keep elevating. Therefore, the numbers of daily active cases are stable and majority of those cases are recovered quickly.

This suggests that the existing businesses should consider partly operating at a ground level.

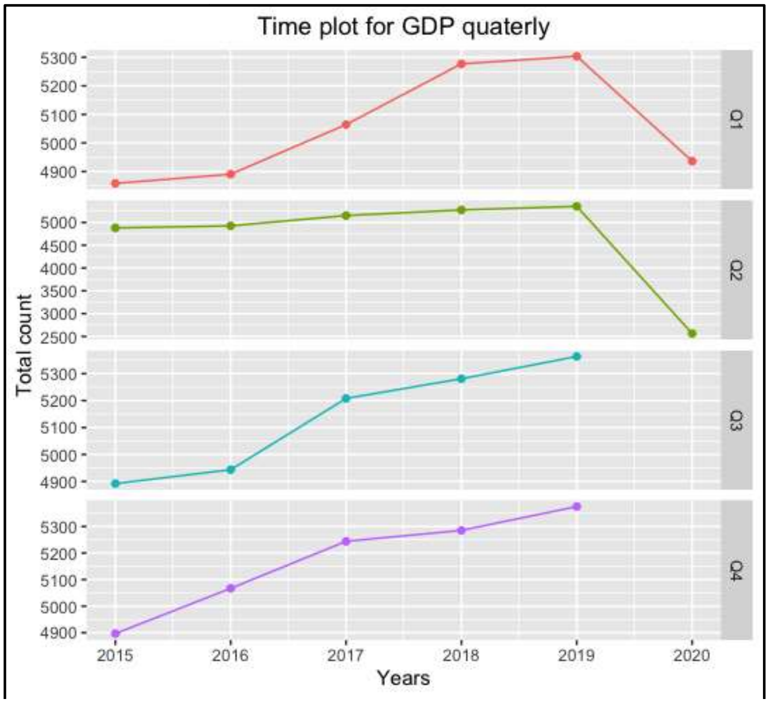


Fig 5

We can analyse the following the above plot:

- There has been a significant rise in global GDP from 2016 to 2019 in all the four quarters.
- In 2020 till now there has been a steep decrease in the global GDP in both quarters which signifies the effect of coronavirus on the GDP
- Overall almost all the countries GDP has by affected due to corona virus.

Then we plotted the graphs of GDP's sum of countries quarterly to show the effect of the corona virus on GDP.

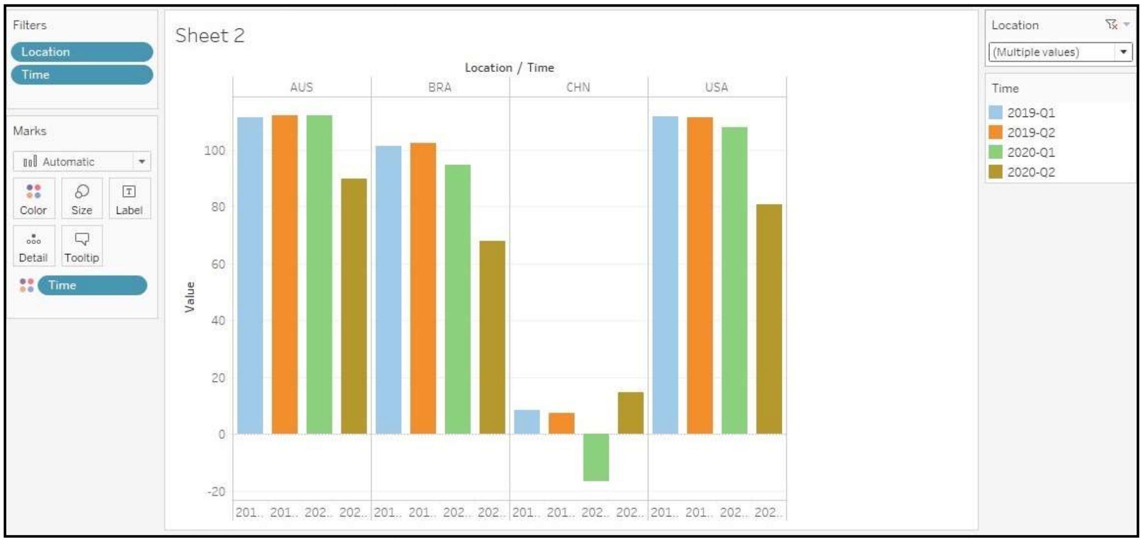
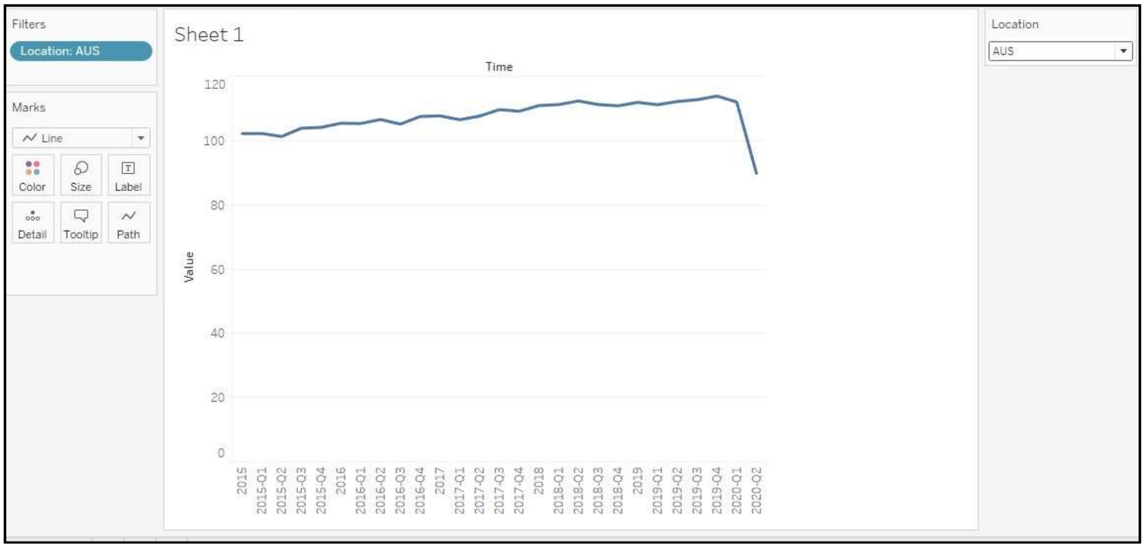


Fig 6

Here we have plotted GDP for 2019 quarter 1, 2 and 2020 quarter 1, 2 for Australia, Brazil, China and USA. We can analyse the following from the plot:

- For Australia the GDP is almost constant for first 2 quarters of 2019 and first quarter of 2020 as the effect of corona virus was not visible in Australia till early march. The lockdown came into effect from April and the GDP of Australia decreased in 2020 second quarter.
- For Brazil the GDP was constant in the first two quarters of 2019 but decreased again in 2020.
- For China the GDP decreased in 2019 first two quarters and then stepped down suddenly in 2020 first quarter (Negative GDP value) showing the effect of corona virus. But later after China recovered from covid and started supplying the equipment's to the world their GDP increased signifying that most of the businesses should move out to china if they want to earn money.
- For USA the changes in GDP is same as that of Brazil constant in 2019 first two quarters and the decreasing in 2020.

We have also plotted time plot for countries using line graph to show the change in GDP from 2015 to 2020



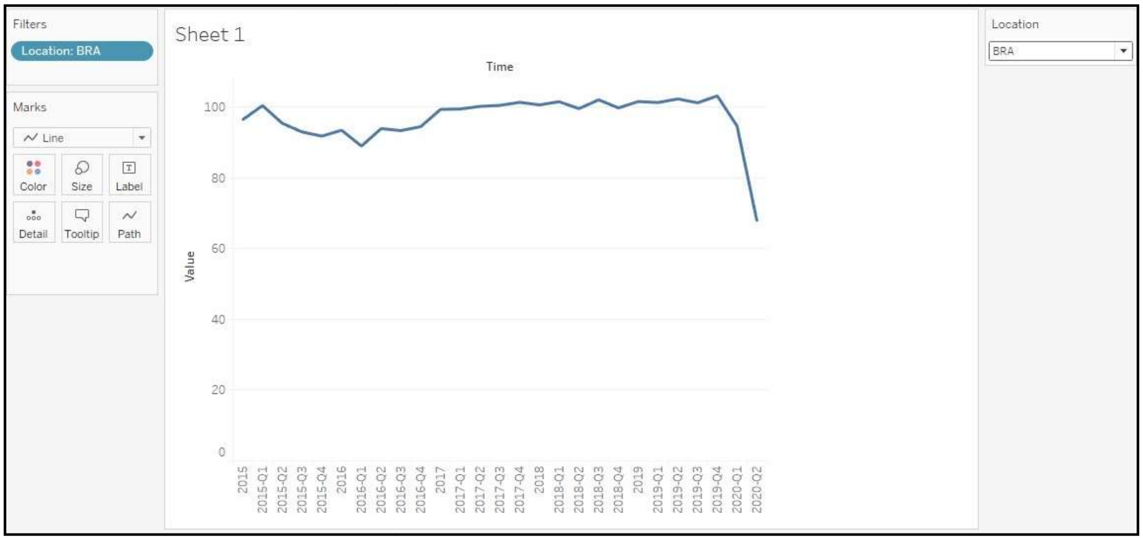
Filters

Location: AUS

Location

AUS

Fig 7



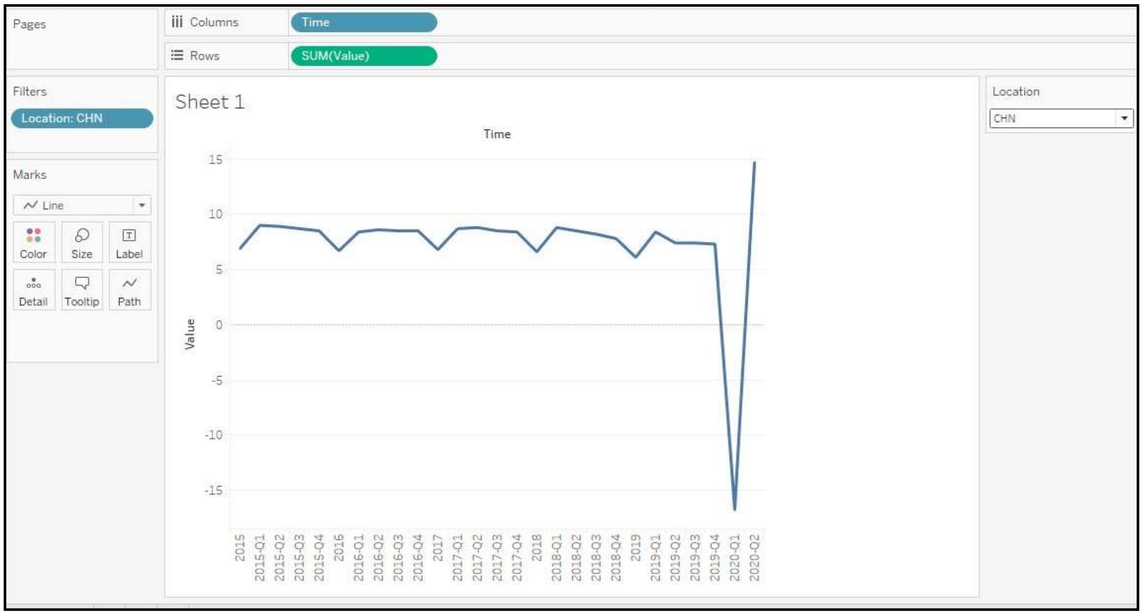
Filters

Location: BRA

Location

BRA

Fig 8



Columns: Time

Rows: SUM(Value)

Filters

Location: CHN

Location

CHN

Fig 9

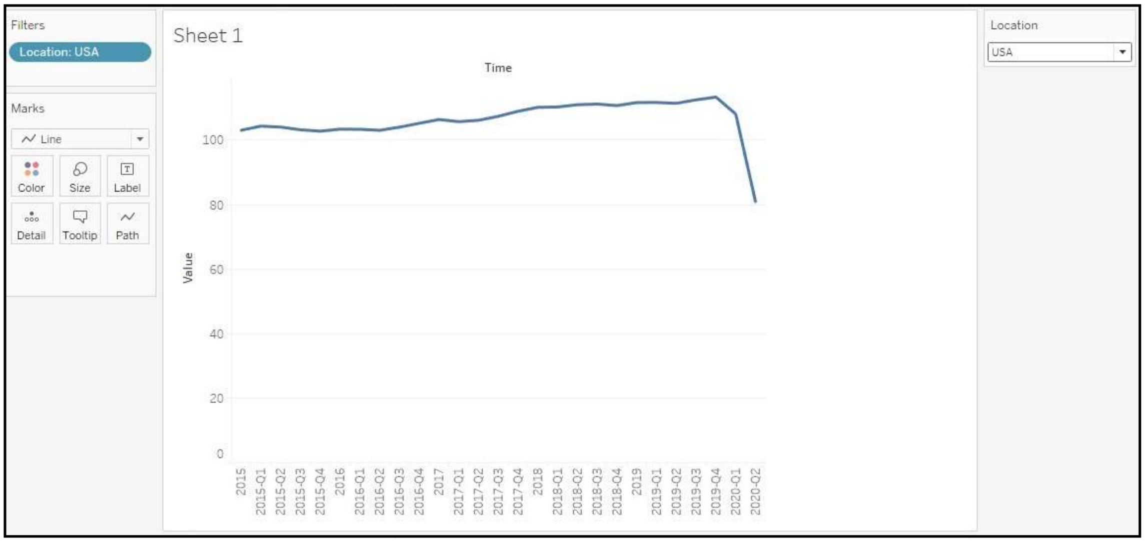


Fig 10

We can analyse the following from the above graphs:

- The GDP for Australia increased from 2015 to 2019 but decreased in 2020 due to corona virus.
- The GDP of Brazil decreased from 2015 to 2016 and increased again from 2017 to 2019 but decreased in 2020 due to corona virus.
- There are significant changes in GDP of china from 2015-2019 as it goes on increasing and decreasing in a constant pattern and then there is as steep decrease in 2020 first quarter. Again, there is a sudden increase in China’s GDP in 2020 second quarter.
- For USA the pattern is similar to that of Australia. The GDP increases from 2015-2019 and then again decreases in 2020.

2.4 Modelling and Validation

For the businesses holding their decision to open up their business operations needed to understand the future GDP in different countries. In order to solve this problem, we built a model which could assist the businesses in selecting countries based on GDP value. The model build could predict the future GDP value for e.g., predicting the GDP of Australia for the next quarter of months.

Since we had already cleaned and balanced the data distribution, now the important part was to build a good accuracy model that could handle any type of data. We had already included the important data attributes (co-relation) by excluding the undesirable data variables in data pre-processing process. Checking the correlation we found that there was high co-relation between variables of data set 1 and likewise for data set 2 variables in merged data set.

As we know, in order to predict something we need to know the behaviour or the pattern of the previous data. So in order to understand the data, training had to be given to the machine and then test the unknown data to predict the values. Therefore, the data was split into 2 sets: one for Train and other for Test. To predict the GDP value based on other features, a *Linear Regression* algorithm was used. This algorithm was used as it works better on numeric values and performs the task to predict a dependent variable value (GDP) based on a given independent variable (other features except GDP).

The data had some categorical data like Country names and Quarters. *Linear Regression* algorithm requires all features to be numeric as it calculates the linear relationship using numeric values. So, those categorical values were encoded to numbers using dummy function. *Linear Regression* was applied on the transformed data and a model was built.

The testing data was fitted into the trained model to predict the values. Subsequently, the results were displayed.

When it comes to numeric prediction there are different measures (prediction accuracy) to *validate* the performance of the prediction model. *Root Mean Square Error (RMSE)* and *R-Squared (R2)* are the measures which were used to validate the accuracy of the model. Moreover, assumptions that are used to test the overall *Linear Regression* model were checked using the outlier, linearity, homoscedasticity, and normalization plots.

The output of RMSE was 3.74 and R2 was 0.57. The RMSE is the difference between values predicted by a model and the actual values observed. The R2 represents the proportion of the variance for a dependent variable that's explained by independent variables. The good model should RMSE close to 0 and in this case it deviates by about 4. This doesn't provide the accurate estimates where generally, GDP values are represented in floats and the decimal difference would matter. Moreover, the R2 value should be close to 1 and this case it's between 0 and 1. Statisticians say that a regression model fits the data well if the differences between the observations and the predicted values are small and unbiased. Unbiased in this context means that the fitted values are not systematically too high or too low anywhere in the observation space [1]. Therefore, we consider the data model is good but not powerful.

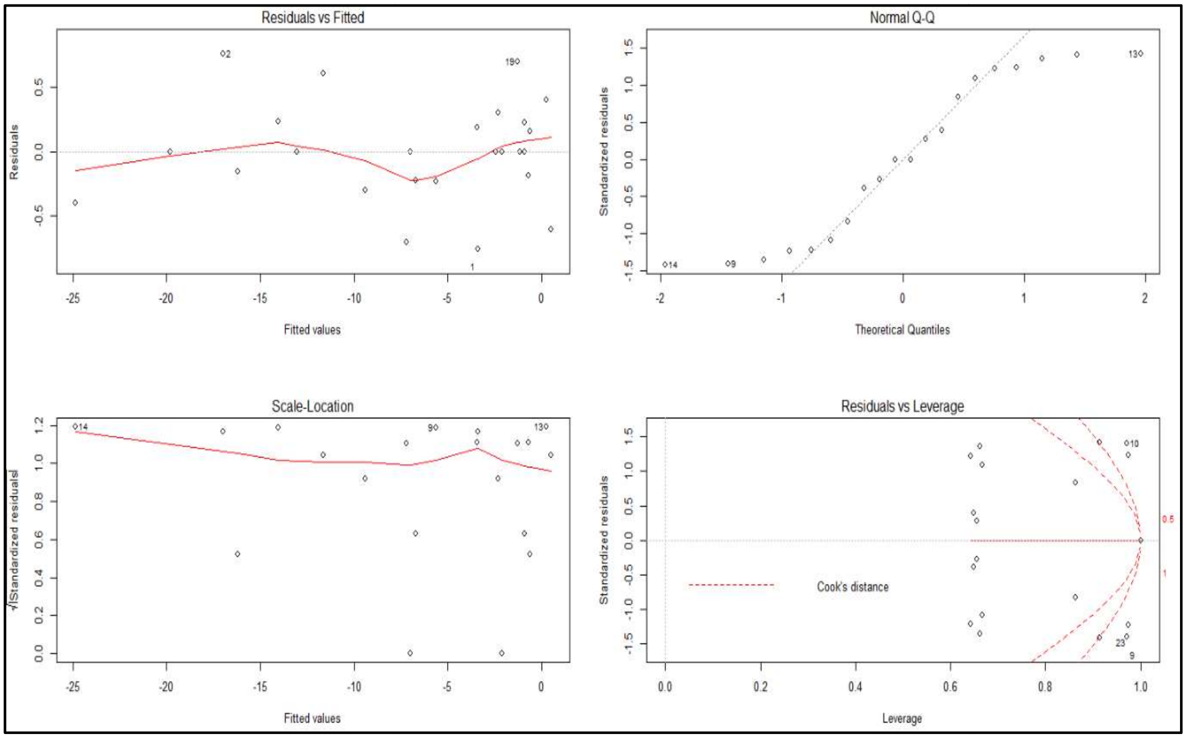


Fig 11

In the Residuals vs Leverage plot (Outliers assumption), the points that are outside the cook's distance are the outliers and here the value is 6. This violates the outliers' assumption but though there are few outliers the model is not affected [9].

The Scale-Location (Homoscedasticity) plot is used to check the equal variance of residuals and a proper equal variance should mean a flat red line. In our case we can see that red line is in the upper area which is almost flat. Although, as we can see the number of fitted values keeps increasing slowly, the gaps between points are higher which should be small [9].

The Normal Q-Q plot (Normalization) is used to check the normal distribution of points where points should be on linear line and shouldn't deviate from it. However, in the Fig 6, we can see light tailed of the points. This also means most of the values are on the extreme side of the values and less no of values on the center side [9].

In Residuals vs fitted plots (Linearity); this plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships. The plot shows that it has a roughly flat line which means that there exists a linear relationship and the linearity is not violated [9].

3 Impact and Significance of Results

The analysis and predictions of the model can add values to the users and can be used by people planning to open a new business in the following ways:

- The dashboard will give user the analysis and predictions of effect of corona virus on the GDP of the country which can help him decide whether to open a business in that country.
- The dashboard will also provide him the insights regarding the corona virus cases and the recovery rate of the country which can help a user estimate about how fast the country can recover from the corona virus.
- Using the insights user can estimate the weightage of safety measures and steps undertaken by the government of the country to stop the corona virus.
- The dashboard can also predict the GDP for the remaining next two quarters using the previous GDP data.
- It can help a user in the decision making of which is the right country to start a business at the moment or also which country is putting the best of the efforts to stop the spread of the corona virus.

4 Conclusion and Next steps

A clearly proposed solution for the problem statement for the business was successfully planned and implemented using the concepts and techniques studied through the course. The data required to reach the proposed solutions was obtained from the John Hopkins University and Organisation of economic corporation and Development. The proposed solution was successfully demonstrated through a working dashboard online on 5th October 2020 through Microsoft Teams. The next steps in the process include the following:

- We aim to incorporate more insights which can help the user make a decision in a better way.
- To expand the model by including the more variables in the dataset which can improve the accuracy of the model and we can get more significant results.
- To make the dashboard and model more accessible to the audience.
- Also, to modify the model in such a way that it can work with the updated data and provide live updates to make business decisions.

5 Project Management Description

First of all, we used a chart in order to track the progress of the tasks involved in each steps of the project. We needed the tools such as Microsoft teams which was required for the effective team communication among all the team members. There were various phase in order to implement this project firstly we needed to do the research phase which means we need to find the appropriate project so we choose covid-19 as our project. Now comes the second phase which was finding the necessary dataset according to the problem statement. We decided the problems which are faced in the real world and those problems were then overcome with our solution and analysis. After that we did all the data type conversions and dealt with the missing values. After pre-processing we have done the analysis and visualizations and gone for the Modelling. Finally, we tested the model and made the final report.

We discussed about the whole project and assigned the list of tasks to the appropriate team members in which they were confident based on their strengths and experience. We discussed every week about the tasks which were associated to the team members. We used trello for assigning task's to each team member and to check the progress of the project. We also planned the next week's activities so that we can do it by next week without any problem. Some tasks took more time to complete while some were completed before time. Each of the team members worked in synergy i.e. if one was lagging behind others helped him to complete the task. We also discussed about the work given to all and also make sure that each member has something to work and give the equal contribution in the project.

We also discussed in the Microsoft teams by taking the meeting after each deadline so that all the members can discuss their tasks and can give solution to the problem if any arises. We also created a WhatsApp group so that we can communicate easily with the team members and discuss the problems. Moreover we used gmail and Google docs for sharing the links and files. Everyone in the group worked very well as a team because all members were very active in the discussions. Everyone in the group made equal contribution to the project.

Below is the table of contributions which were made by each members and also Gantt chart is shown in table 2.

Members	**Contribution**
Nitin Tundwal (s3572060)	20%
Christangel Fargose (s3794800)	20%
Mark Pereira (s3797413)	20%
Shonil Dabreo (s3835204)	20%
Hemanth Kumar Mahendra Kumar (s3823997)	20%

Table 1:- Contributions of each team member.

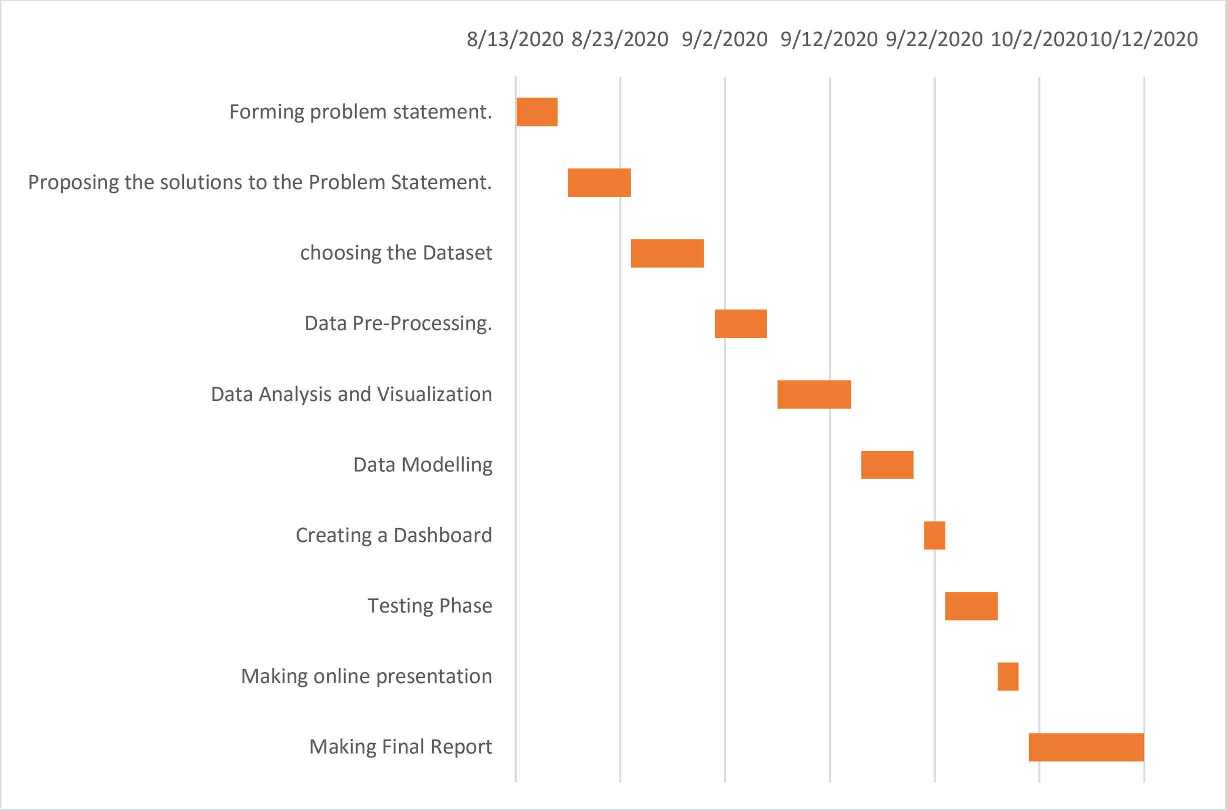


Table 2:- Gantt chart

6 References

[1]J. Frost, "How To Interpret R-squared in Regression Analysis - Statistics By Jim", *Statistics By Jim*, 2020. [Online].
Available: https://statisticsbyjim.com/regression/interpret-r-squared-regression/.
[Accessed: 21- Oct- 2020].

[2]"Impact of COVID-19 on people's livelihoods, their health and our food systems", *Who.int*, 2020. [Online].
Available: https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems.
[Accessed: 21- Oct- 2020].

[3]"Which countries have protected both health and the economy in the pandemic?", *Our World in Data*, 2020. [Online].
Available: https://ourworldindata.org/covid-health-economy.
[Accessed: 21- Oct- 2020].

[4]"CSSEGISandData/COVID-19", GitHub, 2020. [Online].
Available: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.
[Accessed: 21- Oct- 2020].

[5]"Quarterly National Accounts : G20 - Quarterly Growth Rates of GDP in volume", Stats.oecd.org, 2020. [Online].
Available: https://stats.oecd.org/index.aspx?queryid=33940.
[Accessed: 21- Oct- 2020].

[6]"Topic: COVID-19: Impact on the global economy", *Statista*, 2020. [Online].
Available: https://www.statista.com/topics/6139/covid-19-impact-on-the-global-economy/.
[Accessed: 21- Oct- 2020].

[7]*Oecd.org*, 2020. [Online].
Available: https://www.oecd.org/berlin/publikationen/Interim-Economic-Assessment-2-March-2020.pdf.
[Accessed: 21- Oct- 2020].

[8]"RPubs - Analysing the impact of covid on countries", *Rpubs.com*, 2020.
[Online]. Available: https://rpubs.com/NightxWalker/678631.
[Accessed: 21- Oct- 2020].

[9]"Understanding Diagnostic Plots for Linear Regression Analysis | University of Virginia Library Research Data Services + Sciences", *Data.library.virginia.edu*, 2020.
[Online]. Available: https://data.library.virginia.edu/diagnostic-plots/#:~:text=Scale%2DLocation,of%20equal%20variance%20(homoscedasticity).
[Accessed: 21- Oct- 2020].