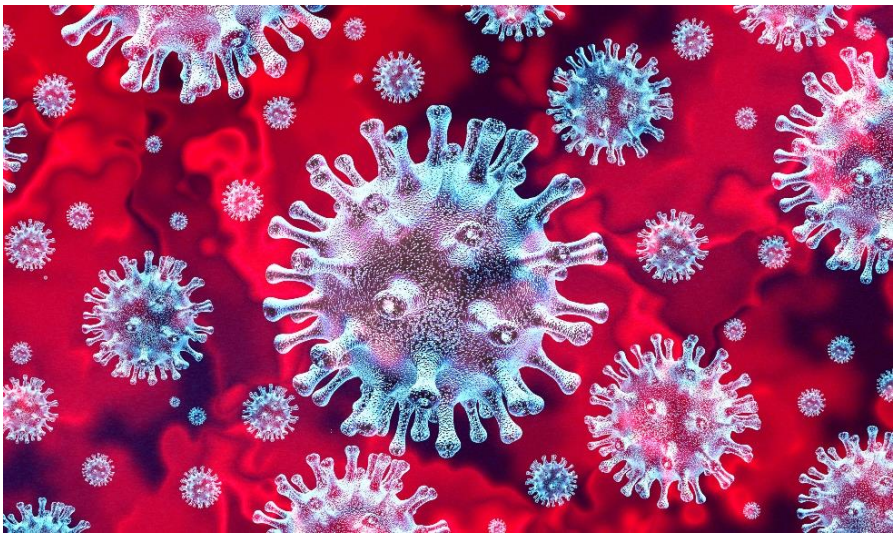
COS3000 – Visualization, Computer Graphics & Data Analysis

Data Visualization Project Report

Name: Calvin John Shaji

Student Number: 45570858

Covid-19: A Comparative Exploration of Spread of The Virus in Australia



Source: https://cerid.uw.edu/image/coronavirus-photo-1jpg

Contents

Introduction

COVID-19 is the disease caused by the coronavirus, SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2). Coronaviruses are a large family of viruses that cause respiratory infections. These can range from the common cold to more serious diseases. The first known case was identified in Wuhan, China, in December 2019. The disease has since spread worldwide, leading to the ongoing COVID-19 pandemic [1,3].

Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop serious illness. Anyone can get sick with COVID-19 and become seriously ill or die at any age. Symptoms include fever, coughing, sore throat, shortness of breath, runny nose, congestion, headache, fatigue, muscle/joint pains, nausea, loss of appetite, diarrhoea, vomiting, temporary loss of smell or altered sense of taste [1,2].

The virus usually spreads from person to person by:

- close contact with an infectious person
- contact with droplets from an infected person's cough or sneeze
- touching objects or surfaces that have droplets from an infected person, and then touching your mouth or face[1].

Aims

This report will examine covid data from January/February 2020 to April 2022 (~802 days), to understand how the *Coronavirus* has spread, as well as comparing Australia's efforts in controlling and minimizing the spread of the virus with other countries. To achieve this, data visualization techniques will be used to find interesting trends in the data and highlight patterns.

During the run of the ongoing COVID-19 pandemic, Australia has enforced several unique rules and guidelines for people to follow to stay safe, with repeated revisions based on the ever-changing circumstances. Over-time, these have been the subject of much controversy with people standing for or against these rules. Examples include the (nearly) 2 yearlong closing of international borders, closing and opening of interstate borders, quarantine requirements, and introduction of numerous lockdowns.

The emergence of different covid variants like the *Omicron* and *Delta* variants, and the introduction of COVID-19 vaccines have also affected the spread of the coronavirus in

Australia. Hence, we want to know how these different factors has affected the spread of the virus in Australia.

Methods

Data Sources, Preparations & Scope

The data used in this report comes from 3 separate sources:

- *Our World in Data (OWID)* website[4,5] (they have sourced their data from *John Hopkins University*)
- *Hum Data (HDX)* website[6] (this is also sourced from *John Hopkins University*)
- *Australian Government Department of Health* website[7]

Due to the immense volume of free resources and data sets available that describe the coronavirus statistics, choosing appropriate sources seemed daunting at first. However, many sources seem to refer to *John Hopkins University*. Due to their reputation on statistics about COVID-19, any sources that refer to the University seemed reliable.

Since we are interested in the spread of coronavirus, datasets that correlate COVID-19 statistics with location seemed appropriate. The major statistics that we are interested in includes the new cases reported per day, total active cases, total confirmed cases, the total deaths, the new deaths per day, and the total confirmed deaths, all based on location (a case* refers to a person diagnosed with COVID-19, deaths* refer to deaths caused by COVID-19). This can give us insights on how much the virus has spread, and the impacts of the spread.

These statistics however were only available from January/February of 2020, they don't include statistics from December 2019. But this is only a small portion of missing data, as we have more than 2 years' worth of COVID-19 data. This is more than capable of illustrating the spread of the virus. The below image is an example of *OWID* data.

	iso_code	continent	location	date	total_cases	new_cases	new_cases_...	total_deaths	new_deaths	new_death...	total_cases...	new_cases_...	new_cases_...
	Categorical	Categorical	Categorical	Datetime	Number	Number	Number	Number	Number	Number	Number	Number	Number
1	iso_code	continent	location	date	total_cases	new_cases	new_cases_...	total_deaths	new_deaths	new_deaths...	total_cases_...	new_cases_...	new_cases_...
2	AFG	Asia	Afghanistan	2020-02-24	5.0	5.0					0.126	0.126	
3	AFG	Asia	Afghanistan	2020-02-25	5.0	0.0					0.126	0.0	
4	AFG	Asia	Afghanistan	2020-02-26	5.0	0.0					0.126	0.0	
5	AFG	Asia	Afghanistan	2020-02-27	5.0	0.0					0.126	0.0	
6	AFG	Asia	Afghanistan	2020-02-28	5.0	0.0					0.126	0.0	
7	AFG	Asia	Afghanistan	2020-02-29	5.0	0.0					0.126	0.0	
8	AFG	Asia	Afghanistan	2020-03-01	5.0	0.0	0.714				0.126	0.0	0.018
9	AFG	Asia	Afghanistan	2020-03-02	5.0	0.0	0.0				0.126	0.0	0.0
10	AFG	Asia	Afghanistan	2020-03-03	5.0	0.0	0.0				0.126	0.0	0.0
11	AFG	Asia	Afghanistan	2020-03-04	5.0	0.0	0.0				0.126	0.0	0.0
12	AFG	Asia	Afghanistan	2020-03-05	5.0	0.0	0.0				0.126	0.0	0.0
13	AFG	Asia	Afghanistan	2020-03-06	5.0	0.0	0.0				0.126	0.0	0.0
14	AFG	Asia	Afghanistan	2020-03-07	8.0	3.0	0.429				0.201	0.075	0.011
15	AFG	Asia	Afghanistan	2020-03-08	8.0	0.0	0.429				0.201	0.0	0.011
16	AFG	Asia	Afghanistan	2020-03-09	8.0	0.0	0.429				0.201	0.0	0.011
17	AFG	Asia	Afghanistan	2020-03-10	8.0	0.0	0.429				0.201	0.0	0.011
18	AFG	Asia	Afghanistan	2020-03-11	11.0	3.0	0.857				0.276	0.075	0.022
19	AFG	Asia	Afghanistan	2020-03-12	11.0	0.0	0.857				0.276	0.0	0.022
20	AFG	Asia	Afghanistan	2020-03-13	11.0	0.0	0.857				0.276	0.0	0.022

The data is a CSV file, with no erroneous data. Data cleansing/processing techniques were not applied to the data, as it was already in a usable format. Plotting some of the data did cause errors however, as some inputs were recorded as NaN (Not a number). These inputs were converted to 0 to work with the data.

The following image is an example of the data from *HDX*, that further correlates new daily cases to a particular location via the addition of corresponding Longitude and Latitude data.

	ProvinceSt...	CountryRe...	Lat	Long	VarName5	VarName6	VarName7	VarName8	VarName9	VarName10
	Text	Categorical	Number	Number	Number	Number	Number	Number	Number	Number
1	Province/St...	Country/Re...	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20
2		Afghanistan	33.93911	67.709953	0	0	0	0	0	0
3		Albania	41.1533	20.1683	0	0	0	0	0	0
4		Algeria	28.0339	1.6596	0	0	0	0	0	0
5		Andorra	42.5063	1.5218	0	0	0	0	0	0
6		Angola	-11.2027	17.8739	0	0	0	0	0	0
7		Antarctica	-71.9499	23.347	0	0	0	0	0	0
8		Antigua an...	17.0608	-61.7964	0	0	0	0	0	0
9		Argentina	-38.4161	-63.6167	0	0	0	0	0	0
10		Armenia	40.0691	45.0382	0	0	0	0	0	0
11	Australian C...	Australia	-35.4735	149.0124	0	0	0	0	0	0
12	New South ...	Australia	-33.8688	151.2093	0	0	0	0	3	4
13	Northern Te...	Australia	-12.4634	130.8456	0	0	0	0	0	0
14	Queensland	Australia	-27.4698	153.0251	0	0	0	0	0	0
15	South Austr...	Australia	-34.9285	138.6007	0	0	0	0	0	0
16	Tasmania	Australia	-42.8821	147.3272	0	0	0	0	0	0
17	Victoria	Australia	-37.8136	144.9631	0	0	0	0	1	1
18	Western Au...	Australia	-31.9505	115.8605	0	0	0	0	0	0
19		Austria	47.5162	14.5501	0	0	0	0	0	0
20		Azerbaijan	40.1431	47.5769	0	0	0	0	0	0

This is also a CSV file, already presented in usable format.

The third data source is from the Australian Government website, and it shows statistics about the cases in Australia and its states, illustrating the source of those cases (i.e., whether it was acquired overseas, locally acquired, and more).

Jurisdiction	Overseas	Locally aquired - contact of confirmed case	Locally acquired - unknown contact*	Under investigation*	Total cases	Total deaths	Locally acquired - interstate travel
Australia	11869	894055	296209	3888365	5113210	6720	22712
ACT	0	91487	0	0	91487	47	0
NSW	4407	165189	1319	1743196	1914104	2588	1312
NT	377	22508	6384	24416	56573	36	2888
QLD	1639	17581	7342	862237	889316	797	517
SA	1942	145137	177265	9222	342673	297	9107
TAS	212	63114	11561	38323	117574	42	4364
VIC	1606	242501	5409	1170567	1420083	2823	0
WA	1686	146538	88248	40404	282400	90	4524

(CSV file)

Methodology

The visualizations were made in *MATLAB* due to familiarity with using its tools, as we have used it extensively to make various visualizations. *MATLAB* is a great visualization tool that makes plotting easy and is also used to perform data analysis. Its basic element is a matrix; hence it is advantageous in plotting data as matrix operations are quick and easy to perform.

Any methods used to perform the visualizations will be discussed throughout the report.

Results & Discussion

Average Cases & Deaths per Day

The first visualizations made were bar graphs of the average reported cases per day, and the average reported deaths per day respectively. These univariate statistics are intended to provide an overall feel of the data, give quick visual insights from the large amount of data available, and illustrate the spread of COVID-19 among different countries.

Initially an attempt was made to make comparisons for all 242 countries available in the data, however, this made the graphs look cluttered and made it difficult to make any meaningful insights from the data. Therefore, changes were made such that the individual bars illustrate these statistics for 20 different countries (mentioned below) including Australia. These countries together make up ~73% of the total confirmed COVID-19 cases today. Excluding Australia, the 19 other countries contribute to ~72% of the total confirmed cases today. The 20 different countries mentioned, along with their country codes are provided below for ease of reference:

United states of America: ***USA****, Great Britain:* ***GBR****, Russia:* ***RUS****, Argentina:* ***ARG****, Turkey:* ***TUR****, Brazil:* ***BRA****, Vietnam:* ***VNM****, Italy:* ***ITA****, Spain:* ***ESP****, France:* ***FRA****, Germany:* ***DEU****, Iran:*

***IRN**, Indonesia: **IDA**, India: **IND**, Japan: **JPN**, South Korea: **KOR**, Netherlands: **NLD**, Columbia: **COL**, Poland: **POL**, Australia: **AUS**.*

An attempt was also made to use histograms instead of bar charts, but their joined bins made it difficult to interpret the data and reduced clarity. The final visualizations of the bar graphs are shown below:

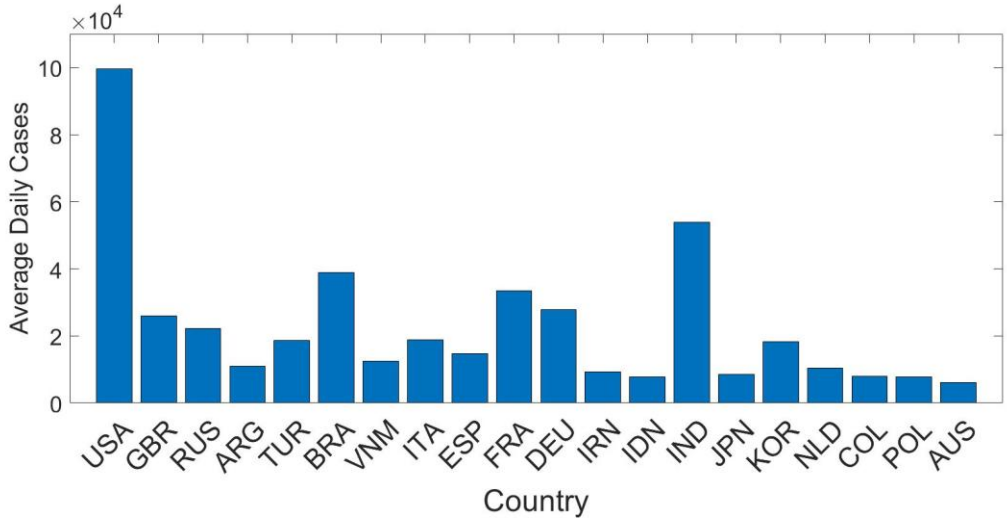


Figure 1 Average daily cases* reported among different countries

Figure 1 shows that USA has the most daily cases on average while Australia has the least daily cases on average. It is also important to see that while USA has the most average daily cases, India and Brazil also contribute to a significant amount of the cases on average.

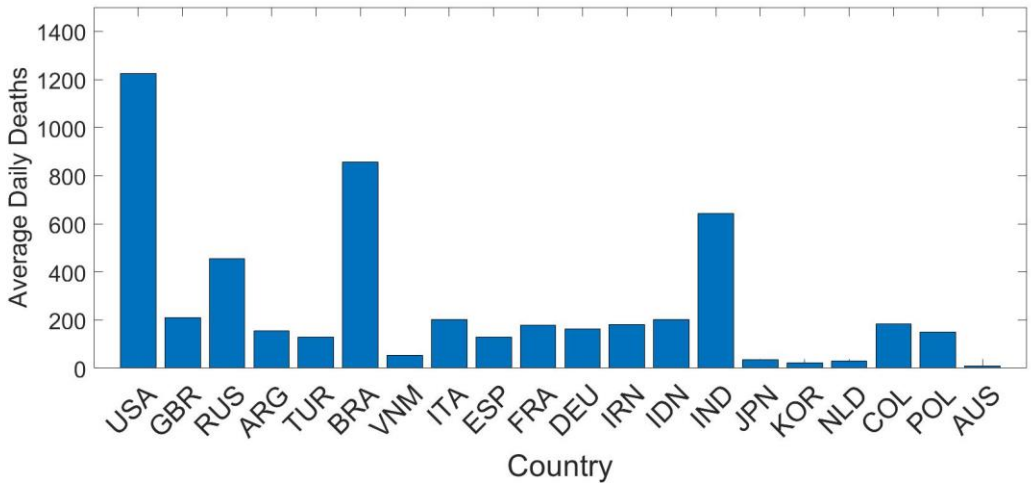


Figure 2 Average daily deaths* reported among different countries

Figure 2 shows that USA has the most daily deaths on average while Australia has the least daily deaths on average. It is interesting to note that USA, Brazil, and India together contribute to most of the average daily deaths, while the remaining countries only contribute towards a much smaller number of deaths on average.

COVID-19 Spread Comparison

The second set of visualizations were bivariate area plots comparing the number of the daily reported cases among different countries. These plots are intended to dive a little deeper into the data, to better understand the spread of COVID-19 cases and compare the statistics of the spread in different countries.

There were some challenges in creating these plots (additional information in appendix). For the purposes of comparison, the plot for each country must be merged to see how they compare. But doing this for 20 countries seemed to be too cumbersome, as overlaying their plots caused confusion due to it being difficult to distinguish colours of each of the 20 plots and identifying the country based on the colour. Hence the comparisons were limited to only 5 countries, namely: USA, India, Brazil, France, and Australia. Together they contribute to ~37% of the total confirmed cases today, and excluding Australia, the other 4 countries are the top 4 countries with the most total confirmed cases.

The plots are shown below:

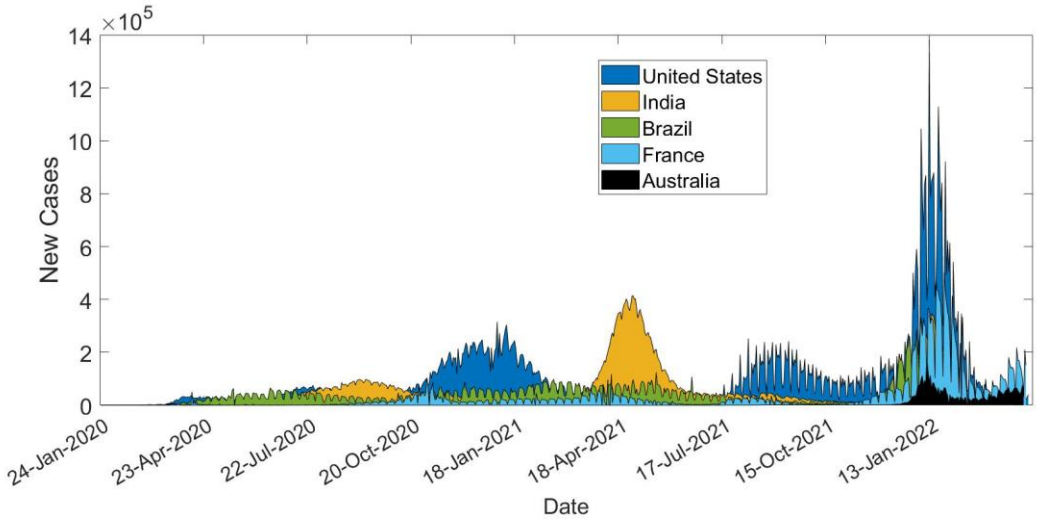


Figure 3 Area plot of daily cases* among different countries

In Figure 3, we see that Australia doesn't seem to have any cases until late December 2021. This is not the case as we will see in the next figure. Australia here consistently has less daily COVID-19 cases compared to the other 4 countries, while USA seems to have much higher spikes in daily cases reported. But what is most interesting to note here is that all the five countries have a massive increase in daily cases as the date reaches the end of December 2021. This is due to the spread of a new variant of COVID-19 at the time called *Omicron*. Omicron is a much more contagious variant of Covid-19, although not as harmful as the variants before. Towards the end of March 2022, the daily cases seem to drop for all countries. Due to the rapid development of COVID-19 vaccines, the spread of Omicron seems to be reducing in all countries in a matter of months.

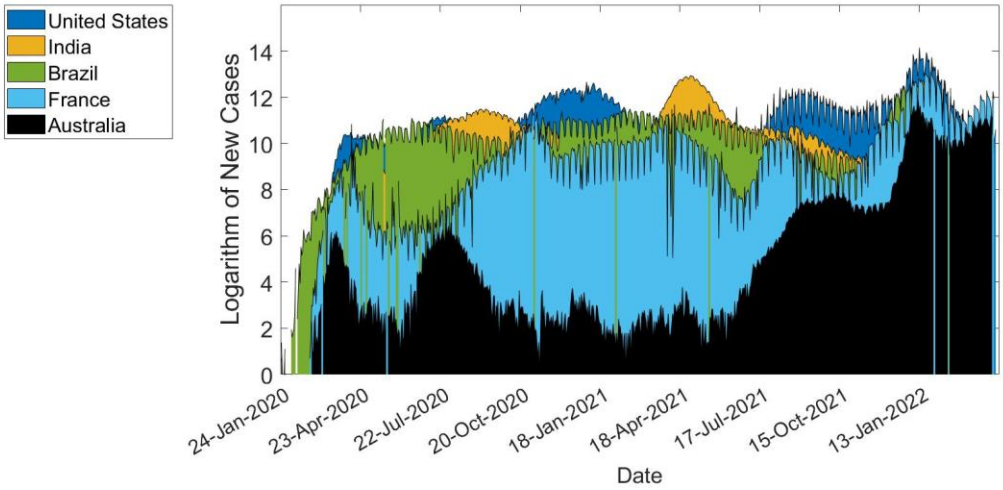


Figure 4 Area plot of the logarithm of daily cases* reported among different countries

Figure 4 shows a better understanding of the spread of cases from a certain perspective. Plotting the logarithm of the daily cases of the different countries shows that the number of reported cases initially was not minimal as one would assume from figure 3. Figure 3 seems to indicate that Australia has no cases until December 2021, which is not true as we can see from figure 4. But since figure 4 is a plot of the logarithm of the daily cases, the differences in plot values along Y axes would mean a much larger difference, i.e., the number of daily cases would deviate much more in reality. We can still see that peak in daily cases for all countries towards late 2021 due to Omicron. It is also much clearer here that Australia does indeed have fewer daily cases on a consistent basis.

A blue/orange colorblind-friendly palette was chosen for the figures 3 and 4 in consideration for readers who may be red/green colorblind. Additionally, black was chosen for Australia in the plots to emphasise Australia, because we are trying to make comparisons of other countries with Australia.

Australia's COVID-19 Spread

The following figures are bar graphs visualizing only Australia's daily reported cases, to focus on the COVID-19 spread in Australia. The aim of these bivariate visualizations is to highlight trends in the relationship between the date and the number of new cases reported. Bar graphs were chosen here because it made the trend easy to visualize.

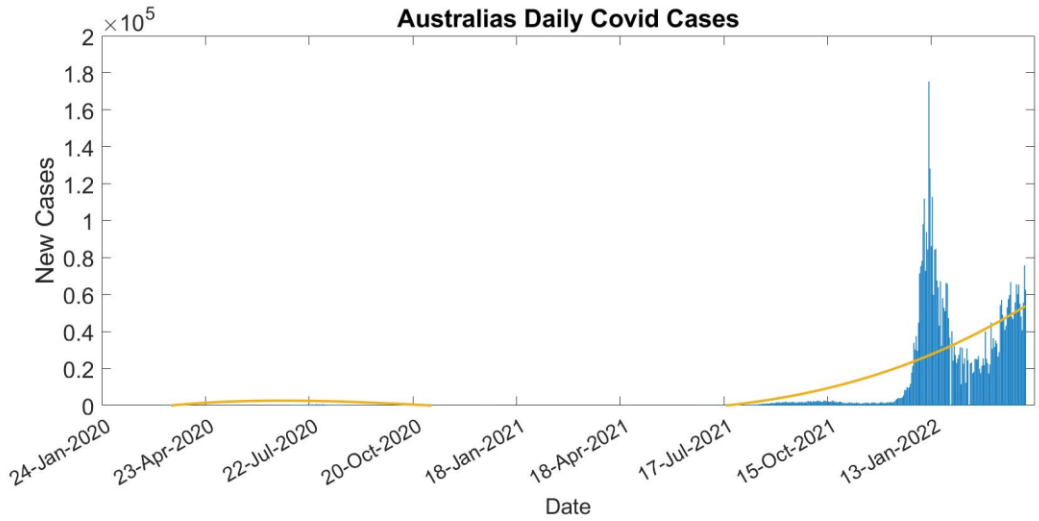


Figure 5 Bar graph of daily cases* reported in Australia

Figure 5 shows that Australia initially didn't have many cases, although it was shown not to be zero from figure 4. Although there was a small increase in cases in the second half of 2021, from January 2020 to late 2021 Australia had very few cases of COVID-19. This is because ever since the virus began to spread, Australia closed its borders completely to non-residents and non-citizens, before completely closing borders for everyone later in 2021. Furthermore, Australia closed its interstate borders for long durations of time, making travel within Australia difficult as well. This greatly limited the spread of COVID-19.

The massive increase in cases late 2021 is due to Omicron, as mentioned prior, but Australia made a controversial decision to gradually open its international and interstate borders at the same time. This was because of the faith in COVID-19 vaccines that were rapidly developing at the time, and COVID-19 was intentionally allowed to spread. There hasn't been a significant decrease in cases since, although there is noticeable decrease.

We can also see that there is an upward trend in the number of cases as indicated by the curve of best fit, but this is greatly influenced by the events discussed in late 2021 and the surge in cases due to Omicron. Hence, this may not be the most appropriate conclusion to make.

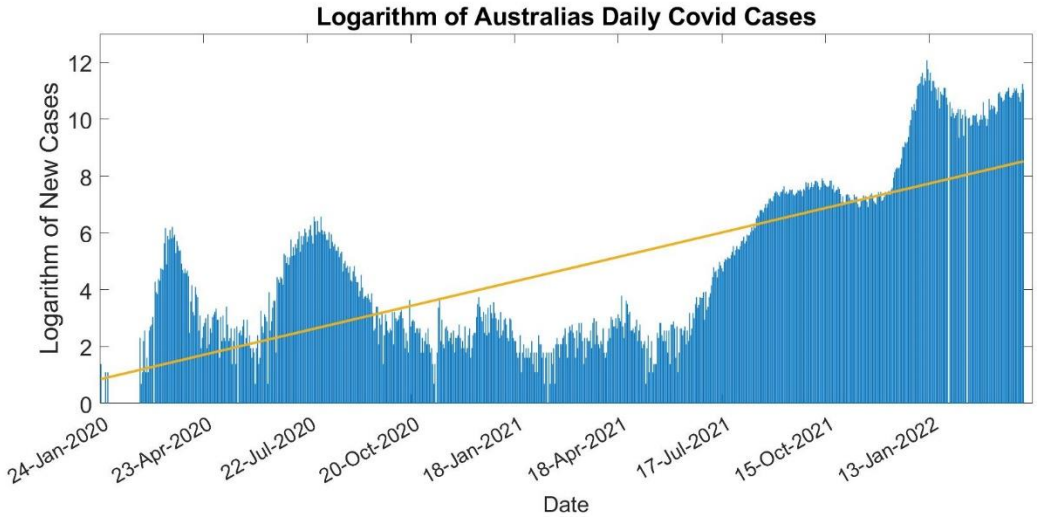


Figure 6 Bar graph of the logarithm of daily cases* reported in Australia

Figure 6 shows the logarithm of the daily cases in Australia, demonstrating that there were more than zero cases in Australia initially. The graph also includes a line of best fit illustrating that cases in Australia continue to rise as time goes on, again this can be inappropriate speculation as the spread in 2022 is greatly influenced by Omicron, introduction of vaccines, and Australian border openings. It is also a logarithm graph of daily cases and does not represent the actual number of cases.

Australia's Covid-19 Sources

To further explore the impacts of international border restrictions, the restrictions on public gatherings within Australia, and other rules that had been put in place restricting mobility within Australia, the transmission sources of COVID-19 cases were explored. Based on the dataset collected, cases were classified based on whether the transmission of COVID-19 occurred via an '*Overseas'* contact, '*Locally acquired – Contact of Confirmed Case'*, '*Locally acquired – Unknown Contact'*, and the rest were '*Under Investigation*'. The following pie chart shows this classification:

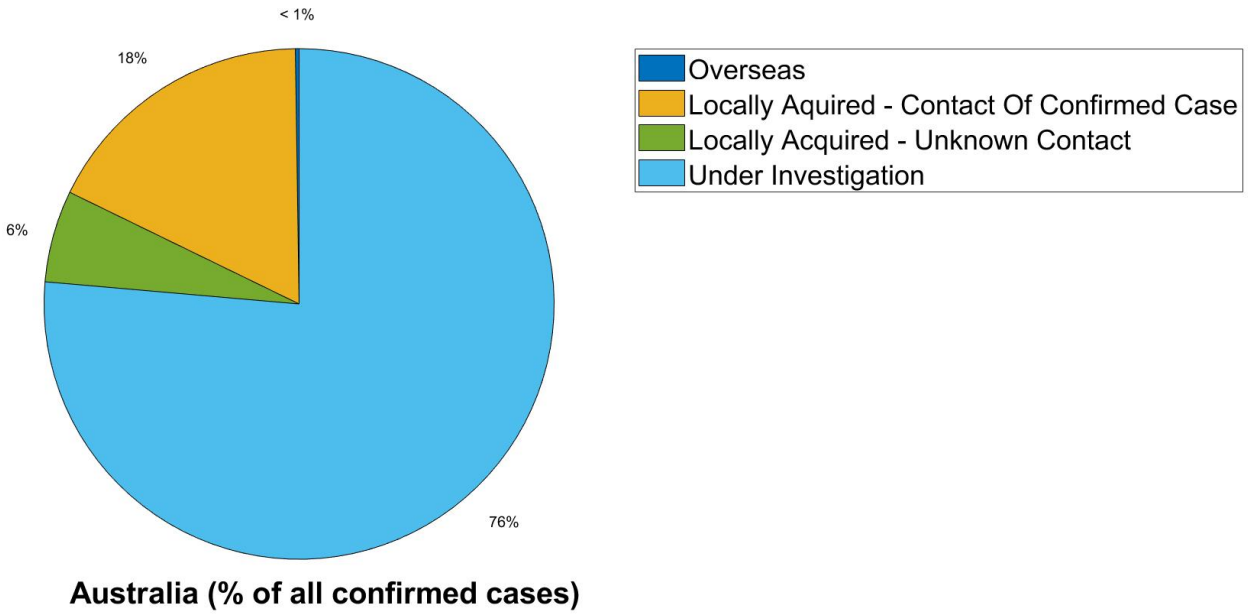


Figure 7 Pie chart of the different transmission sources of confirmed COVID-19 cases in Australia

Figure 7 shows that 76%, i.e., most of the cases are under investigation, meaning we cannot make any inferences for most of the cases. Hence, it would not be a good idea to make conclusions based on the visualizations in Figures 7 and 8 (below).

Less than 1% of the confirmed cases were transmitted via an overseas contact. This is expected as Australia had its borders closed throughout the pandemic, but it may not be reflective of the cases in 2022 as most of them were due to omicron which originated in South Africa (overseas). 24% of the cases were locally acquired within Australia, of which 18% have known contacts.

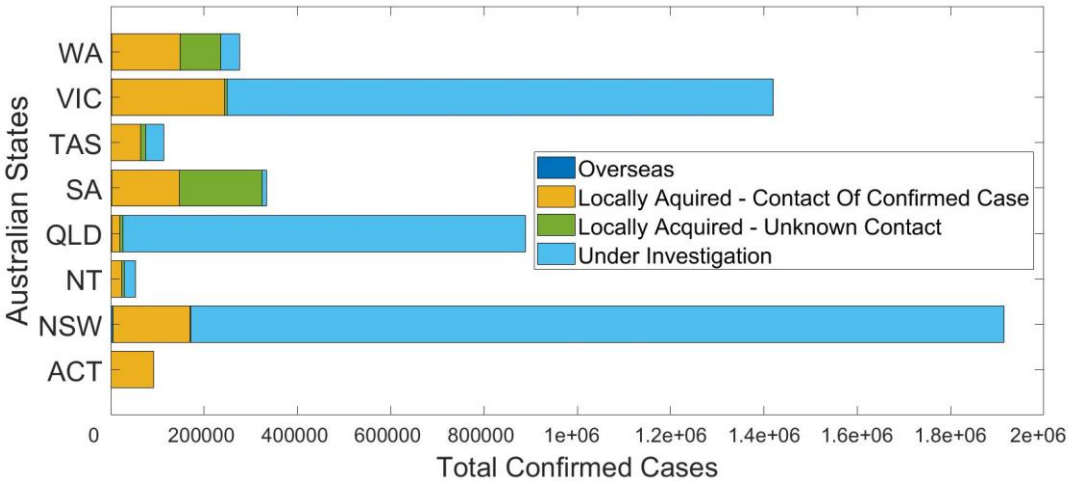


Figure 8 Stacked bar graph showing transmission sources of confirmed COVID-19 cases in each state of Australia

Figure 8 is a stacked bar graph showing the transmission sources of the cases for the individual states within Australia. Again, we see that most of the cases are under investigation. Note, in ACT, all the cases are locally acquired, and contacts are known. SA

and WA also only have a small proportion of their cases under investigation. The cases via overseas transmission is barely visible for all the states as there are very few overseas transmissions identified. It is also interesting to note that while NSW has the most confirmed cases, VIC has more locally acquired cases. Most of the locally acquired cases in NSW and VIC also have identified contacts.

A blue/orange colorblind-friendly palette was chosen for the figures 7 and 8 as well in consideration for readers who may be red/green colorblind.

World Map View of the Spread

The following figure is a World map view of the spread of COVID-19 cases. The visualization was made to give a strong visual overview of the spread of COVID-19, and thereby make comparisons of the spread in different countries with Australia. This was done using MATLAB's inbuilt *geobubble* function.

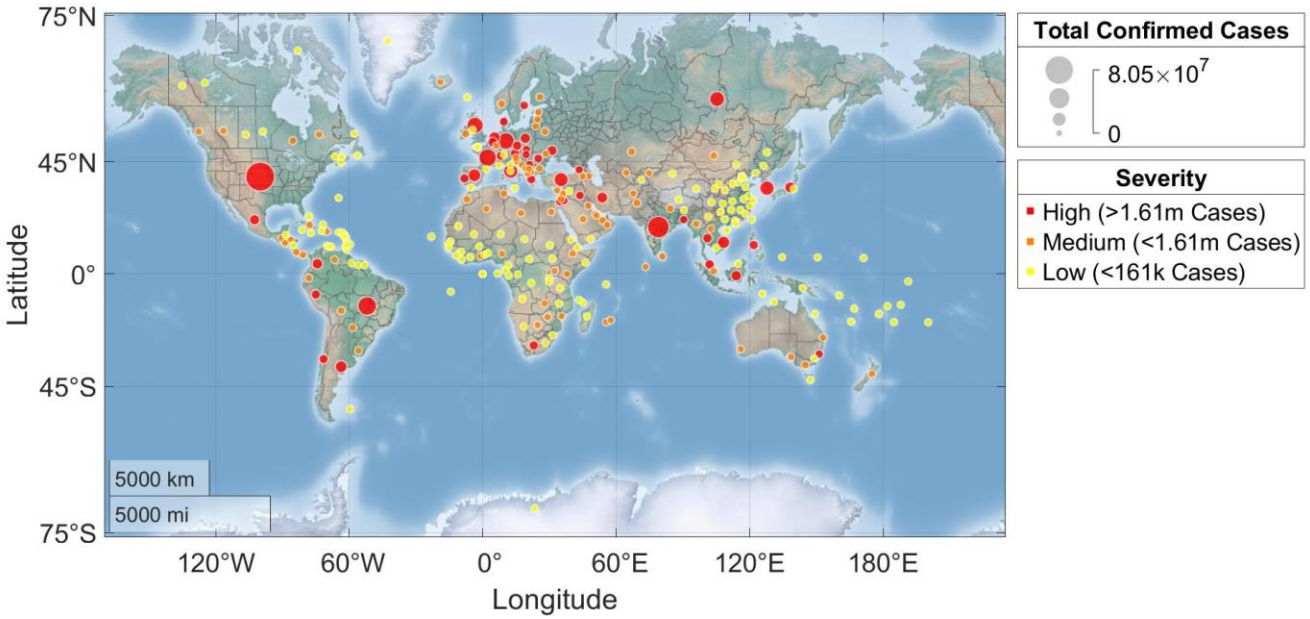


Figure 9 World map view of the spread of covid-19 cases

In Figure 9, the map allows you to distinguish the number of COVID-19 cases in each "bubble" based on its size and colour. The bigger the size of the bubble, the more the COVID-19 cases, spanning a larger area in the map. The colour of the bubble describes the severity of COVID-19 based on the number of cases in the bubble. If there are less than 161,000 cases in the bubble, the severity is low, and the bubble is coloured yellow. If there are more than 161,000 cases but less than 1,610,000, then the severity is medium, and the bubble is coloured orange. If there are more than 1,610,000 cases, then severity is high, and the bubble is coloured red. The numbers classifying the severity were hand-chosen to show a good spread in the bubble colours and enhance the visualization.

The two big red bubbles in USA and India stand out, with more than 80 million and 43 million cases respectively, these regions have been severely impacted by COVID-19. There are also numerous small yellow bubbles scattered in Africa and east Asia near China, indicating that there is a widespread in COVID-19 in these areas, but the severity is low.

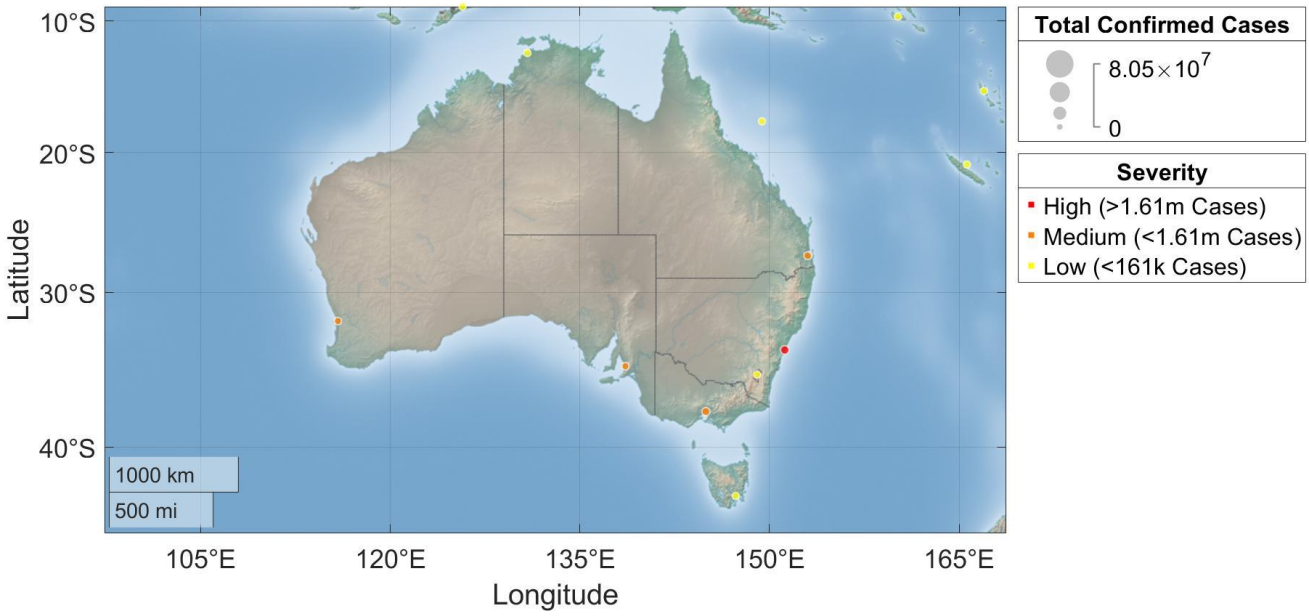


Figure 10 Australia's map view of the spread of covid-19 cases

Focussing in on Australia (Figure 10), we find that there are very few bubbles scattered throughout Australia. This is in part because the Australian population is much denser around its popular cities like Sydney and Melbourne.

We see that there is only one red bubble in Sydney. This is not surprising as we know that there was a rapid spread in COVID-19 cases in New South Wales throughout the pandemic. It was unexpected to see however, that Melbourne had an orange bubble as this was also an area highly impacted by COVID-19. Perth, Adelaide, and Brisbane also had orange bubbles, suggesting a medium severity in COVID-19 cases. Regions like Darwin, Canberra, and Tasmania were not severely impacted by COVID-19.

Conclusion & Limitations

We have compared the daily average COVID-19 cases and deaths in Australia, compared the trend of COVID-19 in Australia, explored factors that correlate with that trend, and explored the geographical spread of the COVID-19 cases worldwide. It is important to note that the visualizations that compared Australia, compared it with countries with the most confirmed cases. Australia may appear to have certain minimal statistics with regards COVID-19, but

there are other countries that we have not compared Australia with that do have lower COVID-19 cases among other statistics.

However, we find that Australia has done relatively well in controlling the spread of the virus, despite the controversy surrounding their actions throughout the pandemic. But this was already apparent. What was interesting to see were the reasons and events that contributed to it. We see how Australia's rules and guidelines (to combat COVID-19), the spread of different COVID-19 variants, and introduction of vaccines have impacted the trend of COVID-19 in Australia by matching the timeline of the events. But causal inferences of the events cannot be made here as these are complex topics that require more research and analysis.

Self-Assessment

I am extremely happy with the variety of visualizations I was able to make with MATLAB. They were relevant, meaningful, and express my practical application of the skills I have learnt. Through the project I was further able to understand my visualizations and the context around it. If I could do the project again, I would have focussed on finding more data about the rules Australia imposed throughout the COVID-19 pandemic and thus create more meaningful visualizations discussing the rules. I would have also liked to research more about the different events that led to the trend in COVID-19 in Australia.

I have avoided basic errors in the report and believe that it is well-written and concise. The report text and visualizations support each other and has achieved the goal of exploring the spread of COVID-19 and making comparisons with the spread in Australia. It shows that Australia has done relatively well in controlling the spread of the virus and illustrates the reasons and events explaining why through useful visualizations. The variety of visualizations included are well-designed and are designed to be viewable by people with colour-blindness (except for figures 9 and 10). Therefore, I believe that this report should obtain a PASS+11 for a grade of a high 6.

Appendix

Figures 10 and 11 below show how figures 3 and 4 were initially plotted (respectively).

The use of regular line plots caused difficulty in making interpretations as the lines were too thin to see in the report. Also, the colours were not easy to distinguish either. Converting the plots to area plots using MATLAB's *areaplot* function in figures 3 and 4 made a huge difference in the clarity of the visualization.

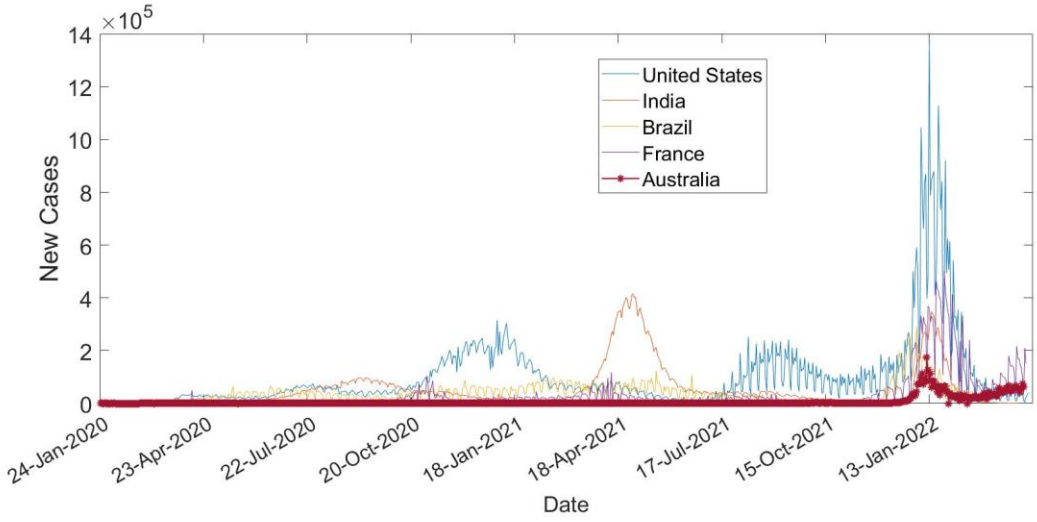


Figure 10 Plot of daily cases* reported among different countries

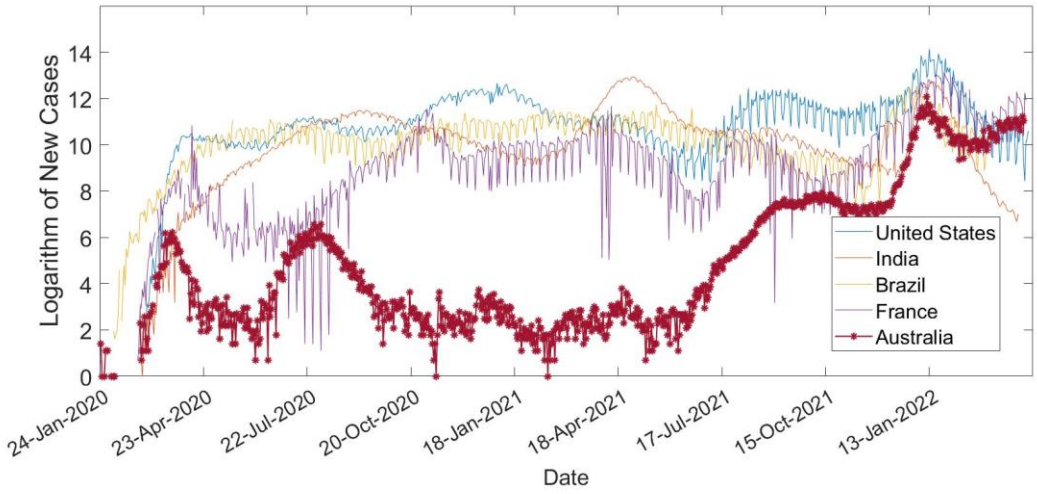


Figure 11 Plot of the logarithm of daily cases* reported among different countries

An attempt was also made at making the area plots translucent, but this reduced clarity in terms of distinguishing colour, and the idea was discarded.

References

[1] “COVID-19 disease, symptoms, and variants.” https://www.health.gov.au/health-alerts/covid-19/symptoms-and-variants, 2021. Accessed: 2022-04-27.

[2] “Coronavirus disease (COVID-19).” https://www.who.int/health-topics/coronavirus#tab=tab_1, 2022. Accessed: 2022-04-27.

[3] “COVID-19.” https://en.wikipedia.org/wiki/COVID-19#:~:text=Severe%20acute%20respiratory%20syndrome%20coronavirus,related%20coronaviruses%20in%20nature, 2022. Accessed: 2022-04-27.

[4] “Coronavirus Source Data.” https://ourworldindata.org/coronavirus-source-data, 2020. Accessed: 2022-04-03.

[5] “covid-19-data.” https://github.com/owid/covid-19-data/tree/master/public/data, 2020. Accessed: 2022-04-03.

[6] “Novel Coronavirus (COVID-19) Cases Data.” https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases, 2020. Accessed: 2022-04-10.

[7] “Coronavirus (COVID-19) case numbers and statistics.” https://www.health.gov.au/health-alerts/covid-19/case-numbers-and-statistics#total-covid19-cases-by-source-of-infection, 2020. Accessed: 2022-04-02.