

Data Exploration & Visualisation Project

COVID-19 Data Analysis

Project Description

Without exaggeration, the COVID-19 infection statistics is currently the most watched and followed information across the world. Given the unprecedented high impact of this epidemic on the health of the world population and the extent of its impact on the financial world, everyone wants to know the scale of the infection across the world but more closely how the local community is impacted.

Visualising the infection data in such a way that it can enable everyone to have information they need is challenging. Moreover, the data is multidimensional with different metrics. The viewer also does not want to be left with a set of graphs. They want clear explanation of them and they want to be able to filter and slice the data the way they wish. They also like to be able to compare different geographic locations. They like to know if daily figures are increasing or decreasing and also how the numbers tally so far. They like to be able to dissect infections in terms of confirmed cases, and those leading to death and recovery.

There is not a single table, map or graph that can answer all these questions. Moreover, there are over 140 countries which makes it impossible to show everything using one or two graphs or tables. What I considered to be the best way to meet these challenges is to provide a newspaper style page that allows the user to start with a global view of the data and then gradually zoom into the details and at the same time empower the user to customise the data visualised by the graph. This style allows us to accompany each graph/table/map with brief and concise description of it to keep the viewer interested.

Where Is the Visualization Published

The html presentation of the visualisation can be found on https://dariush.shinyapps.io/assign3_dariush_v2/ which is publicly available. It does not need any special software of data.

Data Collection & Quality

There are multiple publicly available sources of data, some being copies of one another and some being old. The John Hopkins University provides COVID-19 data for the globe and is one the most recognised sources for this set of data. I have downloaded the following data sets on the 13th April 2020 from this address: <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>

- time_series_covid19_confirmed_global.csv
- time_series_covid19_deaths_global.csv
- time_series_covid19_recovered_global.csv

Having gone through the visualisation of these data sets, I observed the following shortcomings which are not related with the data capturing but the incompleteness of the source of data.:

1. Except for small number of countries such as Australia, Canada and to some extent UK, the data is not available at the state level.
2. In some cases, the cumulative figures are revised down making it difficult to calculate the increment.
3. The number of deaths and recovered is not consistently captured/reported.
4. In some cases, the number of confirmed cases is not being reported on some days.

The data format for all of the three data sets is identical. They have the following format:

- State/Region
- Country
- Longitude
- Latitude
- Cumulative numbers. Each day is captured separately on a different column.

There are four main challenges with these data sets before they can be used for data visualisation:

- There is multiple of them. We would preferable want one data set containing all data.
- The number of cases is not reported in an incremental way. They are cumulative.
- There can be any number of days.
- It is not possible to visualise the data with each measure for each day appearing on a separate column. We want them to appear on multiple rows instead.

Data Wrangling

The aim of the data wrangling is to address the challenges above and arrive at a single csv file by the name: **time_series_covid19_ALL_global.csv** of the following structure:

1. Country
2. State
3. Date
4. Incremental Confirmed Count
5. Incremental Death Count
6. Incremental Recovered Count
7. Cumulative Confirmed Count
8. Cumulative Death Count
9. Cumulative Recovered Count

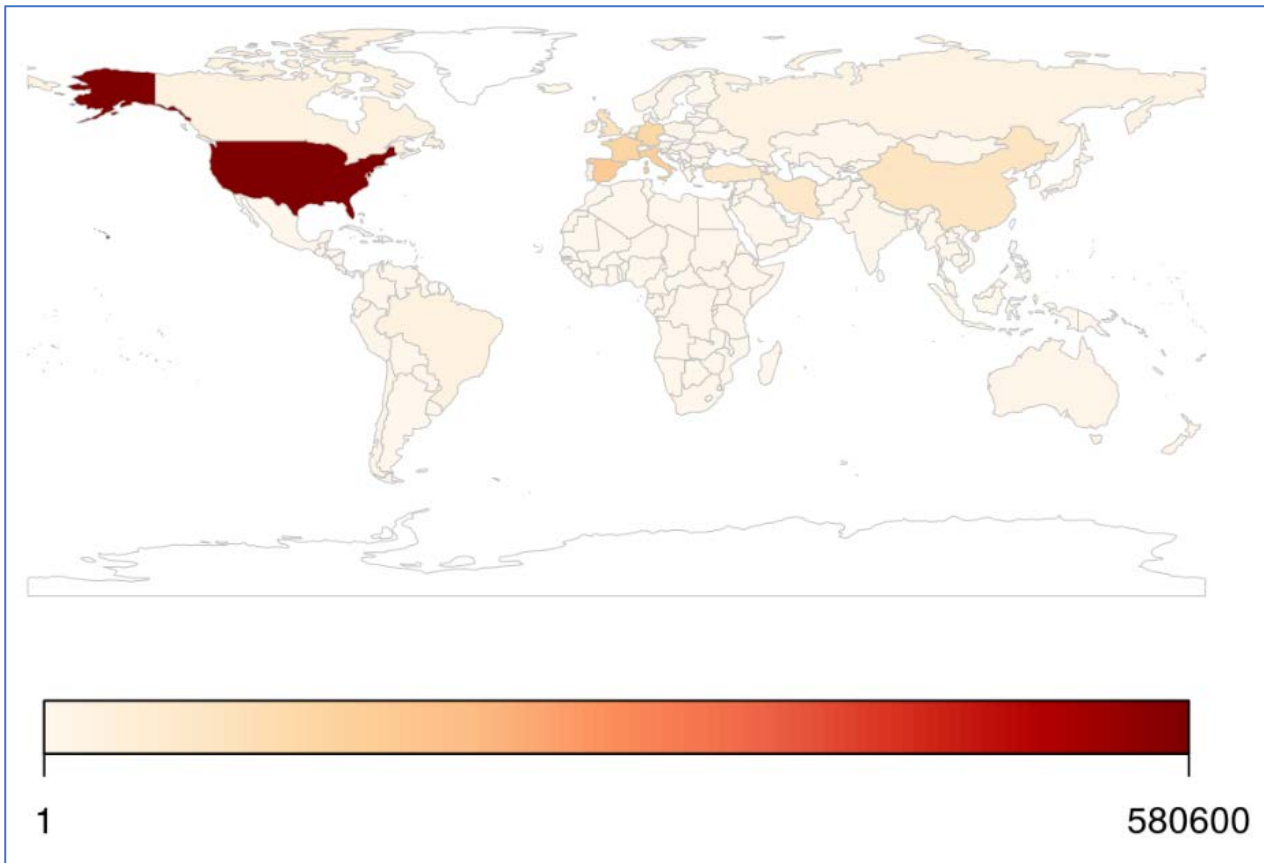
We also need to:

1. Replace missing state with their country name.
2. Set the incremental value to zero if there was a revising down on the figures.

Data Exploration

The layout I have settled with is a newspaper article layout to be able to explain the data as a story. Starting with an abstract, a high-level view and summary and then gradually drilling into areas, categories and metrics. The principal selected is to enable the viewer to dice and slice the data as they wish rather than having lots of graphs. The audience in mind is the general universal public but also additional attention to the Australian case.

Global View of the Confirmed Cases at First Glance



This is a telling picture of where the world stands in terms of the affected nations across the world. We can quickly see that other than in Greenland and South American Nicaragua, the rest of the world is affected. Attention is quickly drawn to the US followed by Europe, then Iran, and China. The colour palette is not precise but provides a relative view of the number of infections. There are other smaller countries such as South Korea but too small to be seen here. When looking at the actual HTML page, it is possible to enlarge the size to get a much clearer picture.

Closer Look at the Global View of All Cases

Having got an idea of the affected nations, a tabular view of the countries with high number of confirmed cases, the following table provides a breakdown of the cases. This table confirms the understanding from the map above but in precise numbers.

- What stands out is that UK and Netherland show very low recovered cases. This ought to be an issue with the reporting as the low figure is not plausible.

- China ranking highest in recovered case.
- What is particularly curious is that recovered cases in Iran is the highest after China even though European nations have a much stronger health care.
- France's case is also curious in that the confirmed cases and death rate is high and recovered cases is very low. If the figures are right, it may be pointing to a serious issue in the management of the crisis.
- Overall, it confirms that Europe followed by the US is the centre of the epidemic.

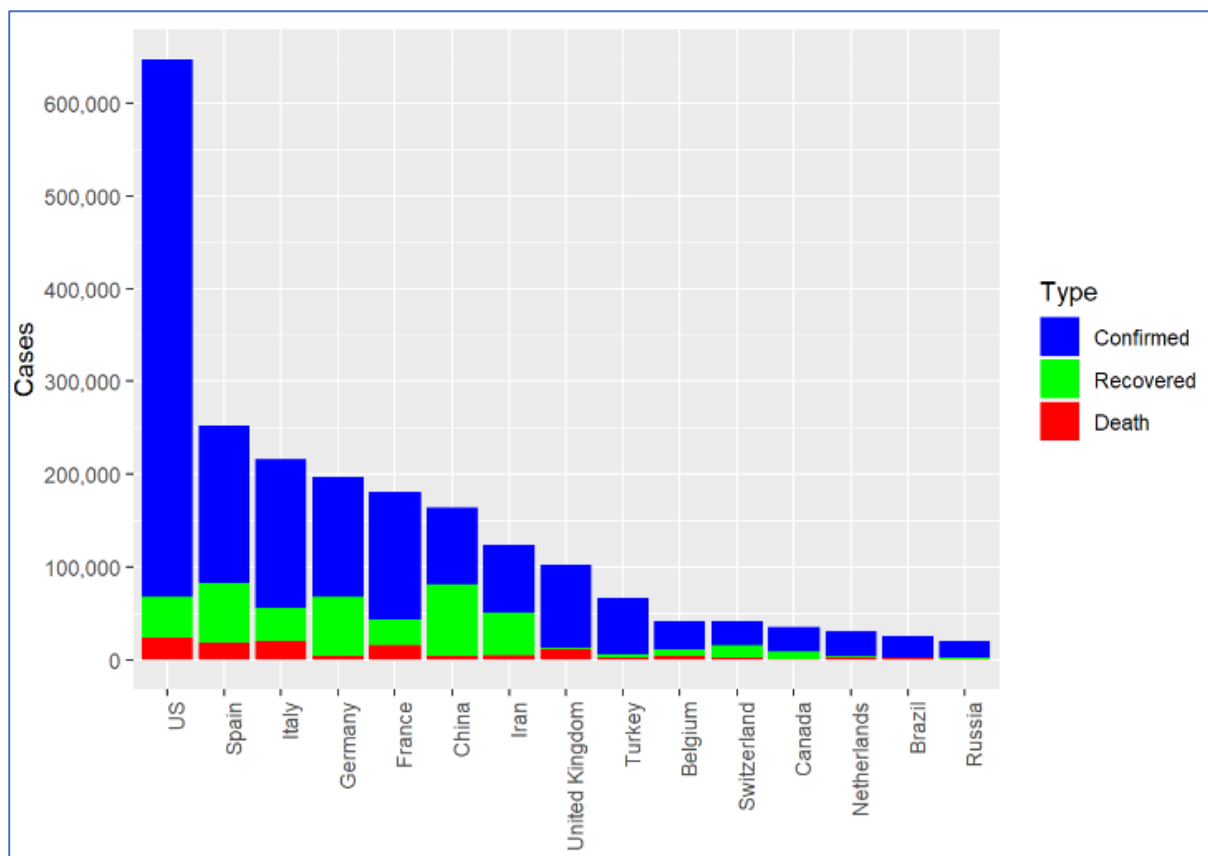
Country	Confirmed Cases	Death	Recovered	Death Ratio	Recovered Ratio
US	580618	23529	43482	0.04	0.07
Spain	170099	17756	64727	0.10	0.38
Italy	159516	20465	35436	0.13	0.22
France	137878	14986	28001	0.11	0.20
Germany	130072	3225	64300	0.02	0.49
United Kingdom	89571	11347	648	0.13	0.01
China	82666	3328	78032	0.04	0.94
Iran	73303	4585	45983	0.06	0.63
Turkey	61049	1296	3957	0.02	0.06
Belgium	30589	3903	6737	0.13	0.22
Netherlands	26710	2833	300	0.11	0.01
Canada	25689	783	8026	0.03	0.31
Switzerland	25688	1138	13700	0.04	0.53
Brazil	23430	1328	173	0.06	0.01
Russia	18328	148	1470	0.01	0.08

The following table shows an aggregation of all cases across the world. It provides a good measure of the global scale of the epidemic. Nearly 2 Million people affected and 120K death with a long way to recovery for people affected.

Total Confirmed Cases	Total Cases of Death	Total Recovered
1916812	119520	449535

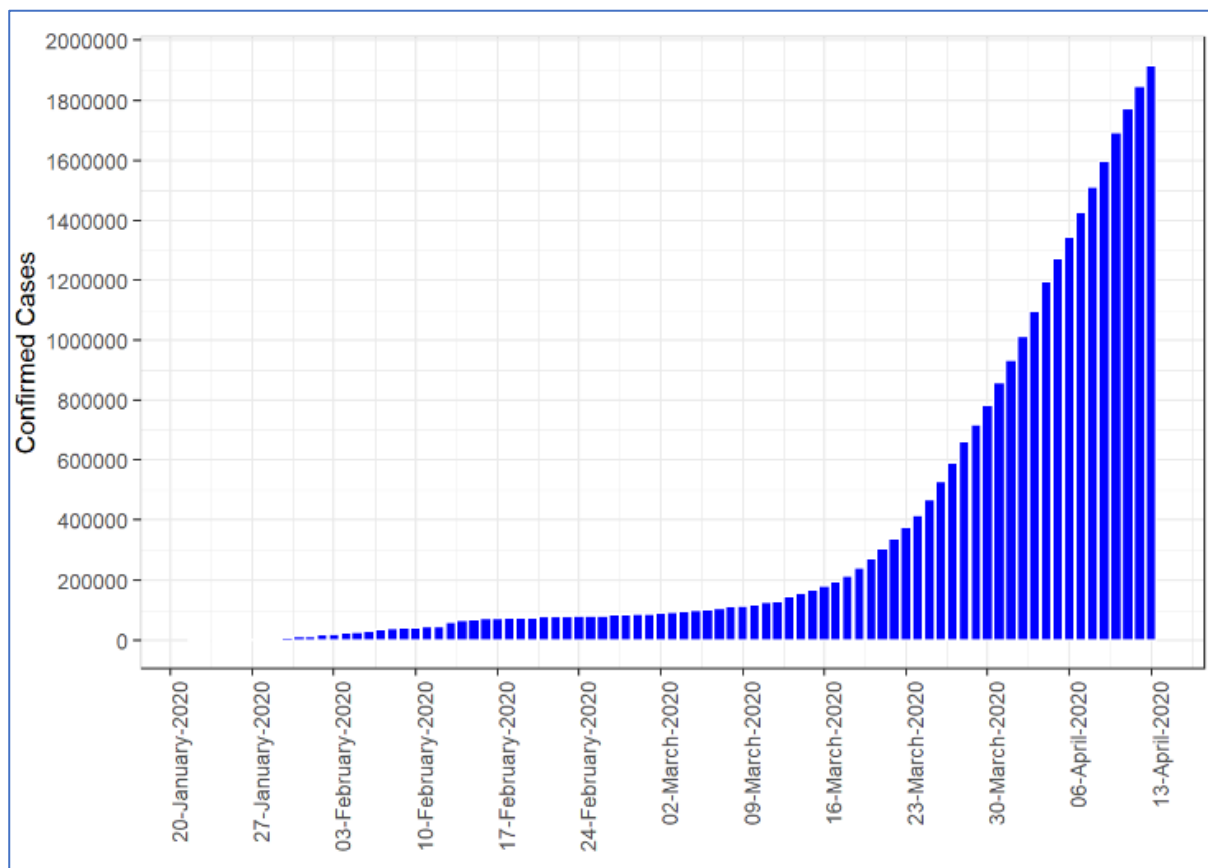
A Visual way of Looking at the Global View of All Cases

The following graph tells a similar story to the table above but in a visual way. It enables viewer to compare countries in each category relative to all other countries. It is more attractive and much easier to make same conclusions here but without much reliance on exact measures.



A Visual way of Looking at the Global View of All Cases

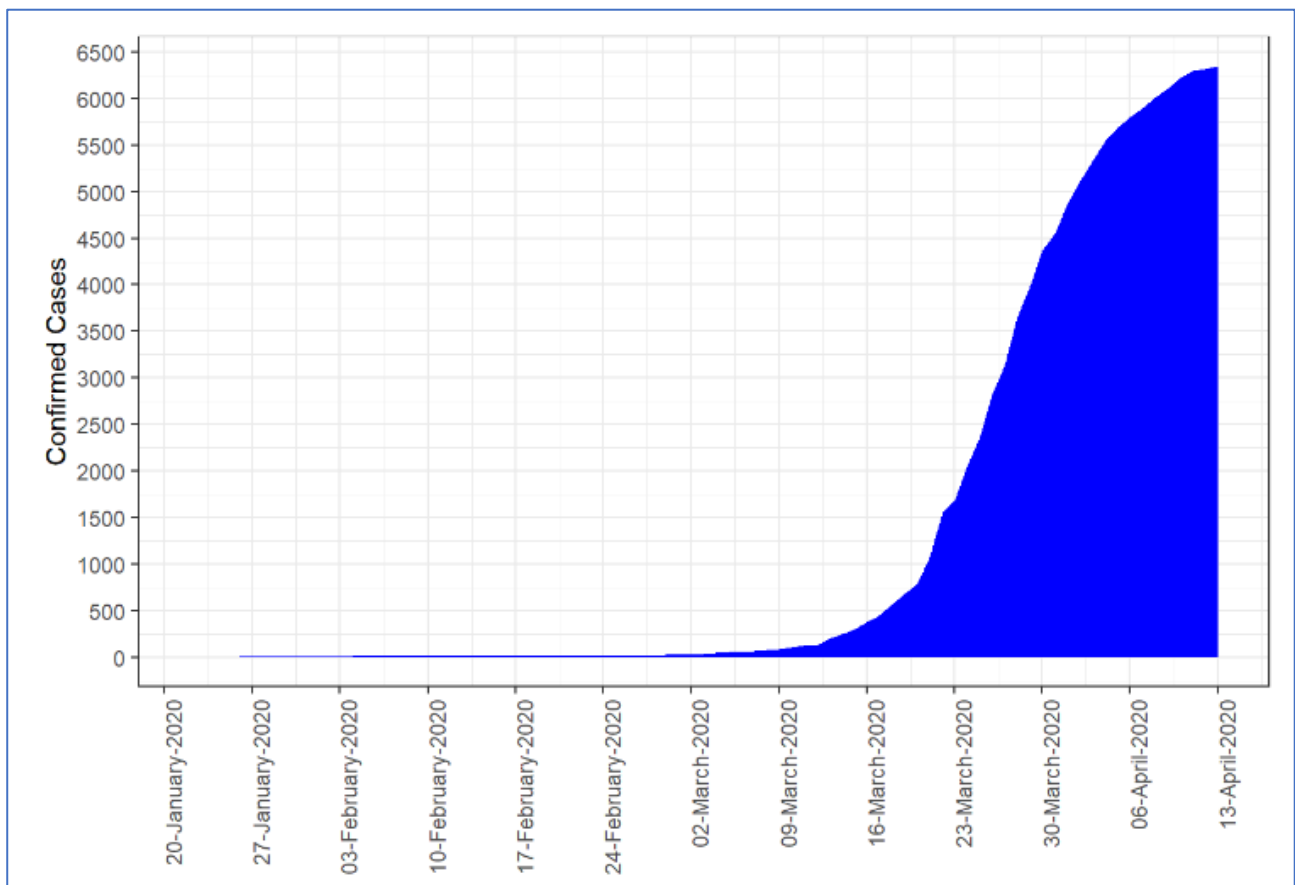
What we are interested to know is how did we get to where we are today?



The above graph shows that the epidemic became more of a global issue around late January with the infections raising steadily but at a very slow pace until mid-March when it started growing exponentially. This milestone was when WHO declared the epidemic as pandemic

How are we Dealing with the Outbreak Here at Home?

We are particularly interested to know how Australia managed the epidemic. Australia was relatively very lightly touched by the virus for the good first two months since the outbreak started. We did not give it much attention until mid-March which also coincides when it was declared pandemic. The **cumulative** number of confirmed cases shown below shows that the numbers were relatively low until early March. Then, the total number of cases started to grow exponentially until late March when it started to stabilize. This was when Australian government started to enforce restriction of movement. It shows that these restrictions were effective in slowing the infection rate. The curve started to flatten in early April.

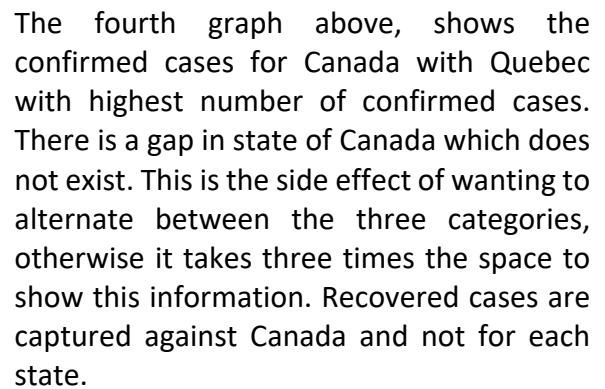


Selective Location Based Analysis of Data for Specific Countries

Whilst study of the trends discussed previously is useful to understand the progression and its rate through the time, we need to know what the current state (i.e. on 13th April 2020) and particularly:

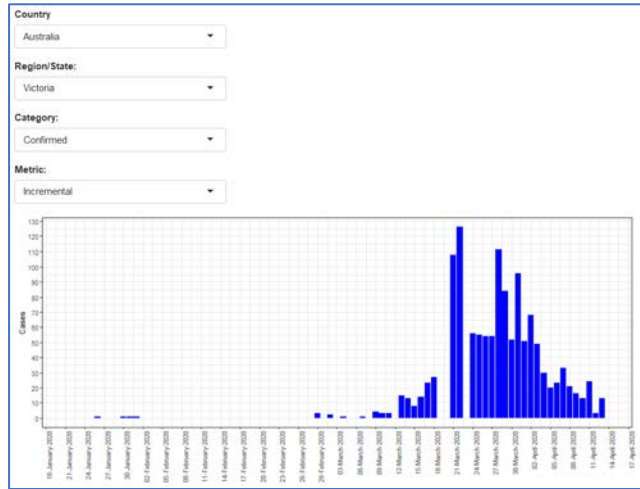
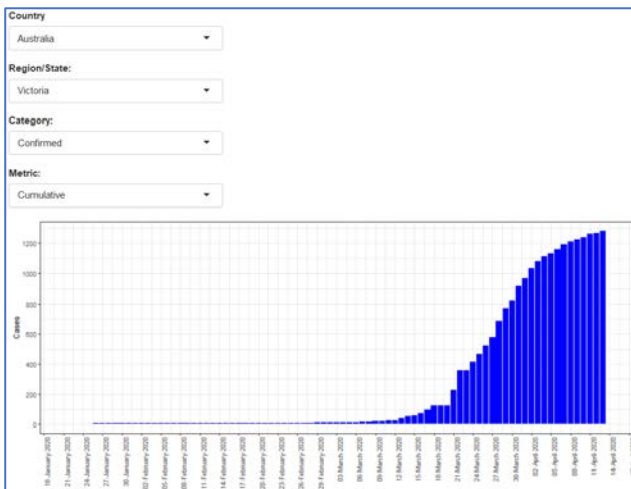
- How are other Australian states doing?
- What is the extent of the death and recovery down to the state level?
- Once we know the above, we also want to know same thing with other countries.

The population difference between NSW and Victoria does not explain the significant gap between the two in terms of infections. The reason must be because of the NSW being more exposed to source of infection from overseas.

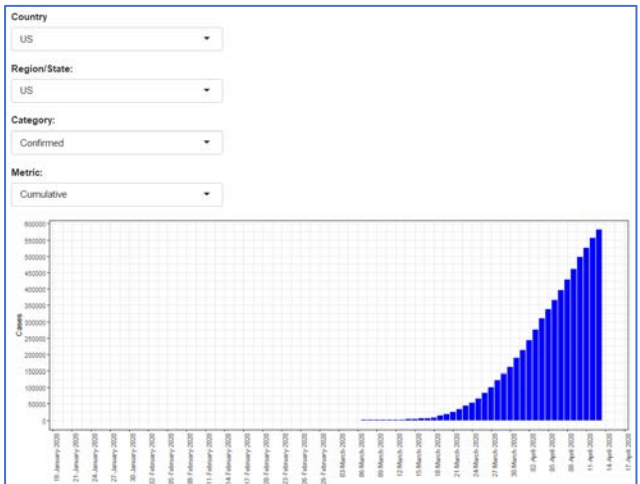
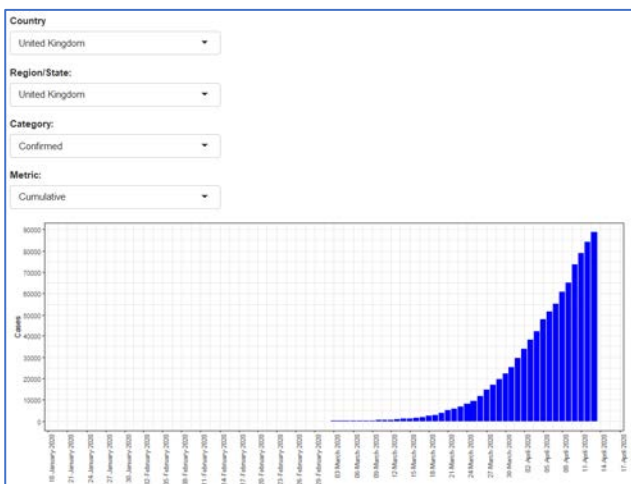
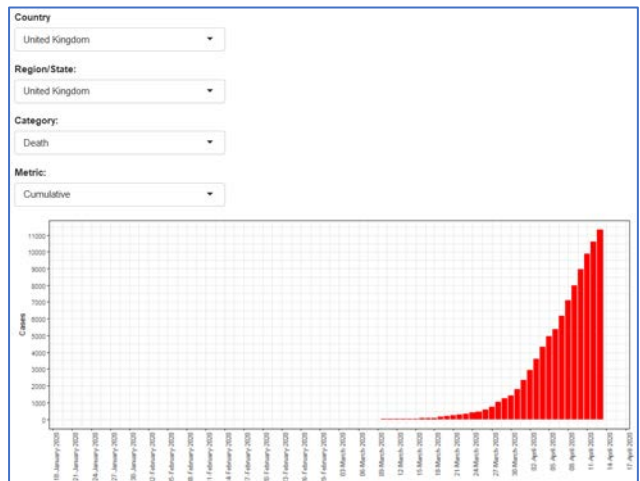
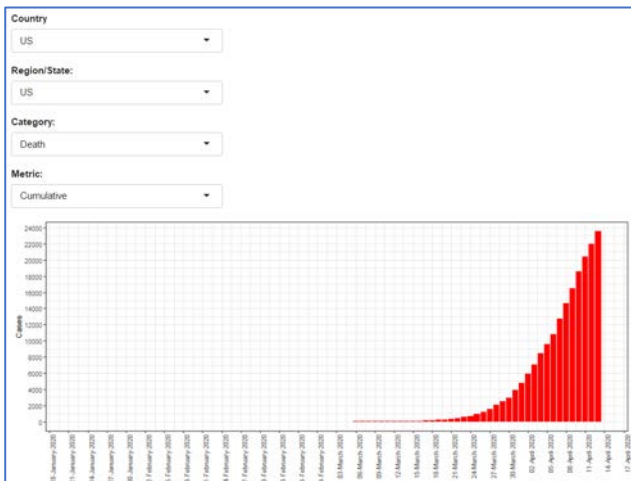


Finally, we like to understand the daily infections as well as cumulative number of cases for each location since the outbreak. To do so, there are 4 different drop downs to be able to drill into each location and examine the incremental and cumulative figures.

The two figures below show incremental and cumulative number of confirmed cases for Victoria. They both indicate that the rate of infection has reduced significantly and at this rate leading to a flattened curve. The number of infections between 5 and 50 vary on a daily basis and there is no visible pattern.



The next two graphs show the rate of death for US and UK with similar pattern in terms of rate of pregression for confirmed and death cases both starting from almost same time pointing perhaps to the travels between the Eurpoe and US which led to US closing its doors to Europe.



Summary

The infection is worldwide and has been widely spread due to the travels. The infection rate when not managed showed to be exponential. Even nations with advanced health care have suffered very high number of cases with large number of cases of death. Restrictions enforced by states have been very effective in reducing the numbers. Whilst number of infections have reduced leading to small daily infections, it is likely that small number of cases will continue to appear for much longer period of time unless it is eradicated with an effective vaccine across the world.

The Design Phase

The design phase evolved out the Five Design Sheet methodology. Having examined the data structure, I realised that I was dealing with a multi variate hierarchical data which is also known as multidimensional data. Although the longitude and latitude of locations were available from the data, it was clear to me that I needed a large area to cover interesting aspects of the data. Therefore I decided to make some high level use of maps.

I came up with the following ideas to choose from:

- A large map to annotate countries with colour, texture and marks to represent the numbers of confirmed, death and recovered cases. I decided that this was difficult to manage specially that I had to show incremental as well as cumulative numbers.
- Tabular representation of data. This was simple and could provide precise measures but would not appeal to the general viewer and would not be attractive. It would be also difficult to draw conclusions from.
- Use of Bar charts which is fairly attractive and provides quantitative measures as well as visual representation. It also enables visual comparison and is not too difficult to implement.
- Use filters in particular drop downs for everything. This is helpful to reduce dimensions but I did not think I could apply to everything because I would lose the overall view and prevents comparison.
- Having examined these alternatives which were also part of the brainstorming of the 5 Design sheet, it became clear to me that I needed to use most of the above but had to balance them carefully. I also gathered that the audience should be the universal public because everyone in the world has very similar questions to which they need answers. I decided for a newspaper article layout in a format (such as html) that could be accessible on the web to the general public.

Using The Five Design Sheets

Sheet-1

I DEAS

* Dimensions of data:
Multi Varate data

Confirmed Cases
Death cases
Recovered Cases

Categories *

Metrics *

Cumulative numbers
Incremental numbers

Global Data
Country
Region
Date

Filter

What in the state of infection globally and locally as well as the trend?
How are the nations coping?

COMBINE & REFINE
HTML PAGE - Article

OPERATIONS

DISCUSSIONS

SUGGESTIONS

Sheet-2
Design-1

TITLE: COVID-19 Data Analysis
Date: 15-APR-2020
Sheet: 2
TASK: Use the 5 sheets to explore

User can get a global view
Darker color suggesting higher infection
Density of + represents Recovered cases
Density of - represents scale of death

Impressed user
- No user operation
- Data is pre-loaded

Color Spectrum
+ Recovered
- Death

Legend
Focus

Confirmed Cases
Recovered Data Set
Deaths Data Set

Focus is on the country and the scale of the infection.

OPERATIONS

DISCUSSIONS

SUGGESTIONS

SHEET-3

State-1
State-2
State-3

TITLE: COVID-19 Data Analysis
Date: 15-APR
SHEET 3

User selects a country
User selects which category to look at
User gets bar charts for each graph
Shows confirmed cases only
Data is filtered based on dropdown selections

Country
Category
confirmed
Death
Recovered

LAYOUT
Focus

OPERATIONS

DISCUSSION

SUGGESTIONS

SHEET-4

Country
DESCRIPTION of Graph

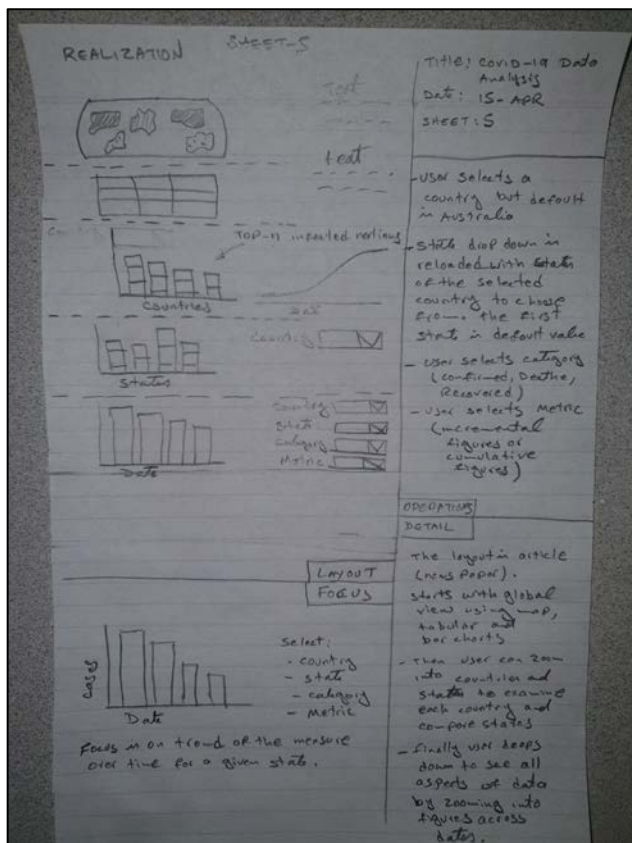
User selects country
There is no date
3 distinct colors are used to show the confirmed, death and recovered cases.
Data is pre-loaded but is filtered to user request.

LAYOUT
Discussion

OPERATIONS

DISCUSSION

SUGGESTIONS



Conclusion

From the visualisations provided, we were able to learn the trend and progression of the outbreak since its beginning across the world and at home here in Australia. We could learn the peak, the daily and cumulative number of cases globally and locally. We learned that some countries despite their better health system were more vulnerable to the virus, indicating perhaps that perhaps less developed nations were more immune to the virus. We learned that the restrictions enforced was our only defence against the disease. We learned from the rapid spread that no nation is immune from an outbreak. Given the number and the movement of people, this virus will continue to contract to a minimum in time but it will not completely disappear.

Implementation

There are two modules implemented to achieve the visualisation:

Assign_3_data_wrangle_version_1.ipynb

This is the data wrangling application as described in the Data Wrangling section. It is developed in iPython Notebook.

Report Operation/Instructions

Running the Data Wrangling

The data wrangling is done through IPython Notebook. The file name is

Assign_3_data_wrangle_version_1.ipynb. It uses standard Python packages pandas, datetime and sys. It uses no particular framework. The reason for using Python is that is widely used for data transformation and batch processing.

Although it is possible to run it from terminal, it is easier to open it from Jupyter Notebook and run all cells. This program makes use of three files:

- time_series_covid19_recovered_global.csv
- time_series_covid19_deaths_global.csv
- time_series_covid19_confirmed_global.csv

These files were downloaded from <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases> on the 13th April 2020. These files are updated on a daily basis. Because they add additional column for each day, you will need to change the end data on line 8 to point to current date:

```
end_date = datetime.datetime(2020, 4, 13)
```

Otherwise, please make sure those files are in the same folder as the program before running it.

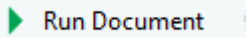
The program generates a new file called `time_series_covid19_ALL_global.csv` in the same working directory. This file is what will be used by the data visualisation process.

Running the Data Visualisation

The data visualisation is written in R Markdown and runs in Shiny mode. Some of the visualisations use Shiny. This program is called `Assign3_dariush_v2.Rmd`.

Packages required are `rworldmap`, `RColorBrewer`, `dplyr`, `knitr`, `ggplot2`, and `scales`.

This program relies on the file `time_series_covid19_ALL_global.csv` created in the previous process. The file and the `.rmd` file need to be in the same folder.

To run the code and see the result, please open the file from RStudio IDE and click on the button  at the top of the screen to run it. It will take around 15 seconds to run. The visualisation appears on the right-hand side.

I have deployed this application on ShinyApps.io under my user name 'dariush'. The report is accessible to all from https://dariush.shinyapps.io/assign3_dariush_v2/. Therefore, you can access the report from here as well.

The choice of using R Markdown was its versatility to create graphs and particularly the use of Shiny to deploy the application as an interactive HTML report. Also, R is widely recognised for data analysis and visualization. One could use Python dashboard or D3 but is more difficult to implement.

How to Access the Visualisation

There are two ways to see the visualisation:

1. Run it through the RStudio IDE as a R Markdown. Please see above for instructions.
2. Access the visualisation from https://dariush.shinyapps.io/assign3_dariush_v2/ as stated above.

Video Presentation

A short view presentation on YouTube is available on <https://youtu.be/OGMmdAhp3Cc>

Conclusion

Whilst there are many excellent visualisations of COVID-19 is available, going through the data analysis myself enabled me to learn key facts about the situation in other countries.

Applying the ggplot2 in practice was particularly useful. Other key learnings were the implementation of Shiny, its integration in R Markdown and its deployments to Shiny Portal.

With respect to Five Design Sheet, although, I went through the process, I am not a great fan of it for 2 reasons: It is an over-kill unless the application is complex and there is a team of people who need to reconcile their design choices. Even so, this can be achieved in many collaborative ways. Agile workshops is a popular method used in the industry.

Links:

- STAT Reports. (March 2020). WHO declares the Corona Virus Outbreak a pandemic. Retrieved from <https://www.statnews.com/2020/03/11/who-declares-the-coronavirus-outbreak-a-pandemic/>
- Shiny Apps. (April 2020). COVID-19 Understanding its Spread and Extent as at 13th April 2020. Retrieved from https://dariush.shinyapps.io/assign3_dariush_v2/
- OCHA Services. (April 2020). Novel Coronavirus (COVID-19) Cases Data. Retrieved from <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>