COVID 19 Data Visualisation using Tableau

Introduction

In our day to day lives, we encounter different and various type of information. This can also vary by size, form and source of obtaining the data. To better understand this, people would leverage on different visualisation techniques. Data visualisation, according to Heitzman (2019), involves gathering of data and converting them into graphical representation to better enhance in terms of presenting insights and meaning about the information. As also discussed by Kelleher and Wagener (2011), visualisation is significant tool in presenting and communicating information as it is able to synthesize huge amount of data into an effective graphical representation. Basically, the purpose of having this can be broadly categorized into two which are data analysis and data presentation.

To be able to perform visualisation, Senay and Ignatius (1994) proposed these steps: Data Manipulation, Visual Mapping and Rendering. On the other hand, Fry (2008) provided more steps to consider particularly in dealing with large data set. These seven steps are shown below.

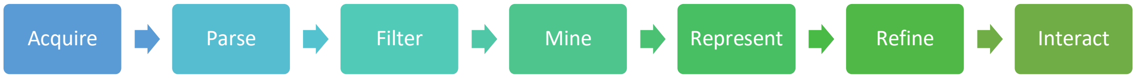


Figure 1: Data Visualisation Steps

The initial step is, of course, to obtain the data from a certain source. This is basically the data acquisition part. After that data parsing would need to be performed. This can mean exploring the data and organize the data structure. Once the structure is organized, filtering can easily be done to ensure that only data of interest are selected. Next would be data mining or for other they also consider this as data enrichment wherein some additional techniques are implemented to be able to understand any pattern that exists. From there, tools and techniques can be identified on how to better illustrate the data e.g. map, bar, scatter plot etc. That can be enhanced by implementing graphic design methods. One way is by ensuring that colour and shapes are properly used to provide better readability. Then finally, interaction can also be added to enable used to explore the data and/or control the features the user wants to see (Fry, 2008).

Tableau as Data Visualisation Tool

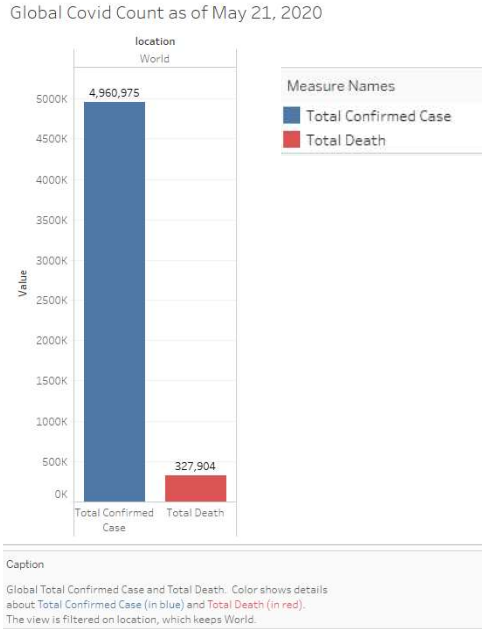
One of the visualisation tools that can be used to implement those steps discussed is Tableau. Tableau, as introduced its website, is said to be a powerful, flexible and secure data visual analytics platform (Tableau, n.d.). Among its various software available, Tableau Desktop, an easy to use interface, will be used in this paper to implement different visualisation techniques. Aside from that, Dataflair (2019) also mentioned the other benefits of using this tool such as high quality of visual image capabilities, strong and reliable performance, multiple information source connection among others. In the succeeding parts of this paper the different visualisation generated from Tableau will be discussed including on how the visualisation steps mentioned were considered.

COVID 19 Data Set

As mentioned earlier, the first part of Data visualisation is to identify the source. In this case, this paper will use the COVID 19 data set from Our World in Data website (Ritchie, 2020). However, it will only consider the data reported until May 21, 2020. As also mentioned in their website. There are four main metrics in data set which are total confirmed cases, total deaths, new confirmed cases, and new deaths. Details on number of tests, population, location, date are also other fields available in the csv source file. Using the VLOOKUP method from Excel, the author has also added new field which is continent wherein the aim is to improve the visualisation that will be performed.

Visualisation Implementation

After exploring the data set, it would be a nice start to show the global count using a bar chart. On the graph below, it shows there the global summary of total confirmed case and total deaths as of May 21, 2020. Note that Total test is empty hence author simply showed these two to provide a global count for this pandemic.



Caption

Global Total Confirmed Case and Total Death. Color shows details about Total Confirmed Case (in blue) and Total Death (in red). The view is filtered on location, which keeps World.

Figure 2: Global COVID Count of Confirmed Case and Death as of May 21

To track the count for both over time, a line chart can provide more details about this as shown below. This line chart is able to provide information regarding the increasing number of cases of both metrics. On below, dual axis was used to plot the on the same graph. It was also helpful to have unique colour on each metric and include a label of the peak of new cases for both. For example, on the week of April 6, there were 52, 531 new death cases which was at its peak. On the other hand, it was on the week of May 11 that the number of new cases was reported to be at its peak of 608,652. Also, another good thing that we can observed on both metrics is the drop on new cases which hopefully would be continuously going down.

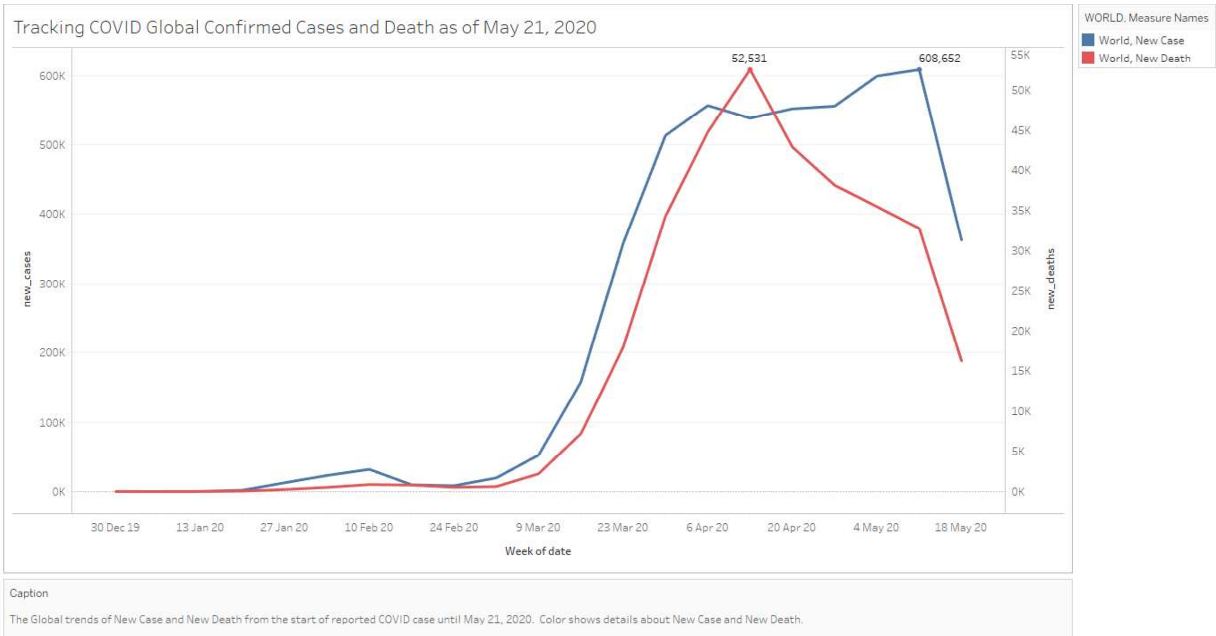
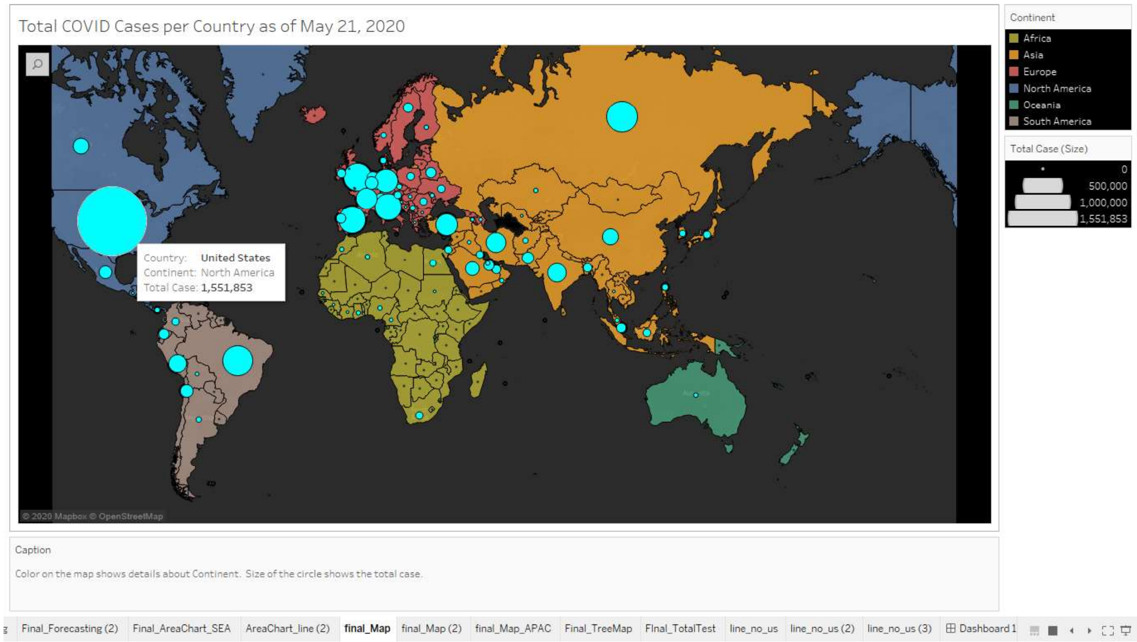


Figure 3: Global Covid Trend of New Confirmed Case and Death as of May 21

From both bar and line graph, it was shown some significant numbers on COVID confirmed cases and death count. Kiss (2018) discussed line graph would definitely be useful to present trend which indeed serve its purpose in this case; hence, it was effective to visualise the global overview. However, if all of the figures of each countries were included it would not be that effective and would not give a clear picture of the information. As mentioned by Gulbis (2016) and ROM (n.d.), bar and line graphs may not be that suitable for large scale visualisation.

Knowing that limitation, another technique that can be explored to somehow show the highlevel information per country is using a map. As mentioned by Gulbis (2016), map can provide an overview (but not an absolute count) of the data across various location in this case country. As also suggested by the same author, using overlay bubble can better improved the visualisation which is implemented in this by adding the blue-green circle. From that, its size can aid in identifying which countries have higher number of confirmed cases. From Figure 4, it can be observed that United States has the most number of cases. By hovering the pointer on the circle, the total case can be checked which is 1,551,853.

Furthermore, each continent was given unique colours. In doing so, another insight can be observed from the map in terms of which continent has more countries having high confirmed COVID cases. From the below map, it can be noticed that there are a lot of countries in Europe with high cases of COVID as shown by the blue-green circles.


Total COVID Cases per Country as of May 21, 2020

Country: United States
Continent: North America
Total Case: 1,551,853

Caption

Color on the map shows details about Continent. Size of the circle shows the total case.

Figure 4: Total COVID Cases per Country using Symbol Map.

The good thing on symbol map as discussed by Mace (2014) is that it can provide additional insights with regards to a location in the map. Shapes and color can help in showing another level of detail of detail which were discussed earlier on how these are relfected on Figure 4. However, another challenge on using this visualisation is avoiding overlapping symbols especially on small countries.

Another way, to illustrate the number of cases per country is via tree map wherein in figure 6 it is partitioned by continent. In this case that overlapping of the overlay bubble from the map can be addressed by tree map. Creating a hierarchy region in dimensions as shown on Figure 5 was necessary to show that partition explicitly.

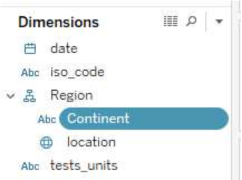
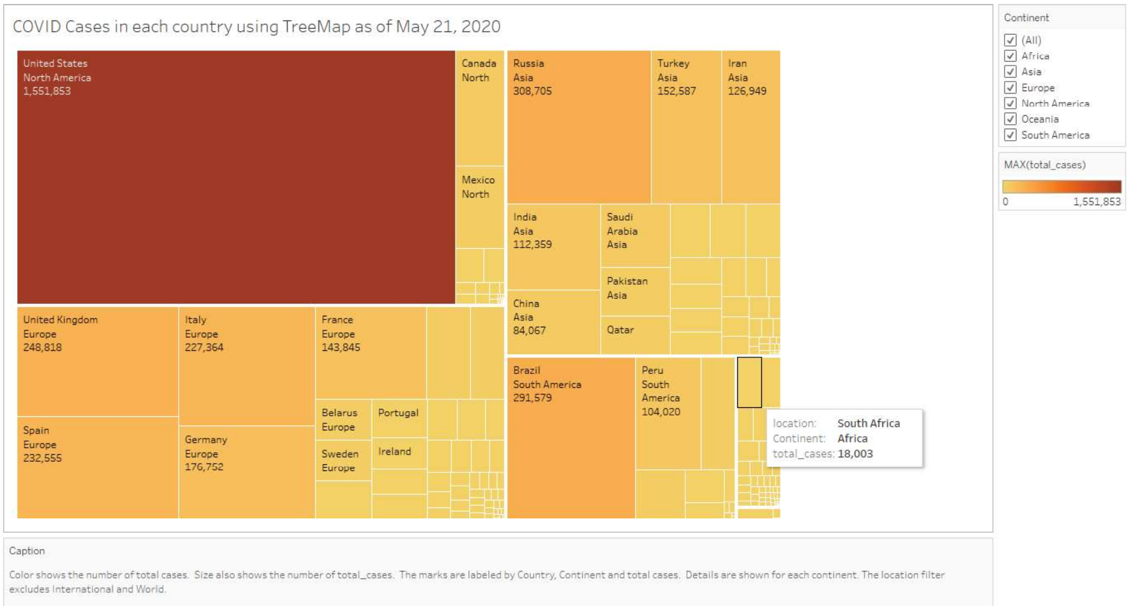


Figure 5: Hierarchy Region



location: South Africa
Continent: Africa
total_cases: 18,003

Caption

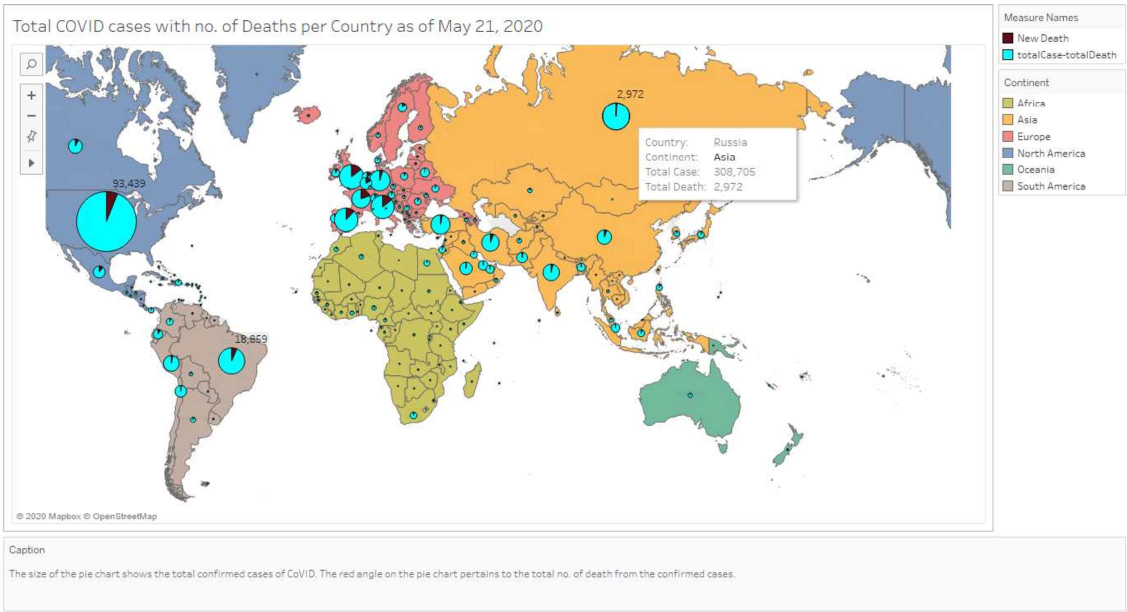
Color shows the number of total cases. Size also shows the number of total_cases. The marks are labeled by Country, Continent and total cases. Details are shown for each continent. The location filter excludes International and World.

Figure 6: COVID Case using Tree Map

From Figure 6, it can be seen that there that indeed US has the highest number of cases. It also intensified the previous observation from the map that there are a lot of European countries with high number of cases.

Shneiderman (1992) mentioned that tree map was initially created to be able to visualize large of amount of relational data. Though it may not explicitly show the relationship it would be able to get insights and patterns at varying level (Plaisant et al. 2004) (Stasko et al. 2000). Another issue in using tree map is that in printed form, it would be of a challenged to clearly know the attributes of the smaller regions with even labels not visible (Shaffer 2017).

Going back to the map in tableau, additional attribute can be shown by integrating pie-chart in tree map. In this case, total death can also be shown. For this, the author decided to create another measure to calculate the total case less death count. Basically, the whole pie in Figure 7 refers to the total confirmed case wherein the light green portion is newly calculated field and the red part refers to the number of death case. In this case, it was able to accurately show the numbers in the pie chart integrated in the map. It can be observed that there are countries with high cases of COVID but have lower reported case of death like in the case of Russia wherein there are only roughly 3K total death as compared to the 308K total cases. On the other hand, Brazil in South America and United States in North America have 18,869 and 93,439 cases of reported death as indicated using marker.

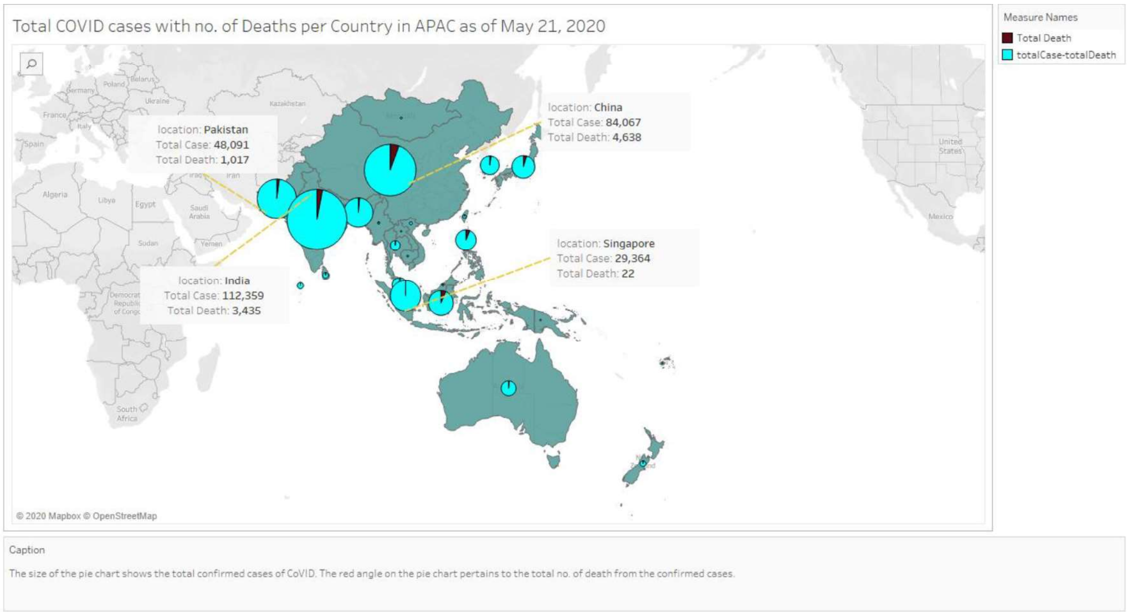


Caption

The size of the pie chart shows the total confirmed cases of CoVID. The red angle on the pie chart pertains to the total no. of death from the confirmed cases.

Figure 7: Total COVID Cases including Death count via pie chart per Country

Another thing to be able to look more into the distribution on some parts of the global is by doing some filter. In this case, it would give us a better picture of the pie chart in some countries. For example, if the region of interest is APAC below figure can provide information.



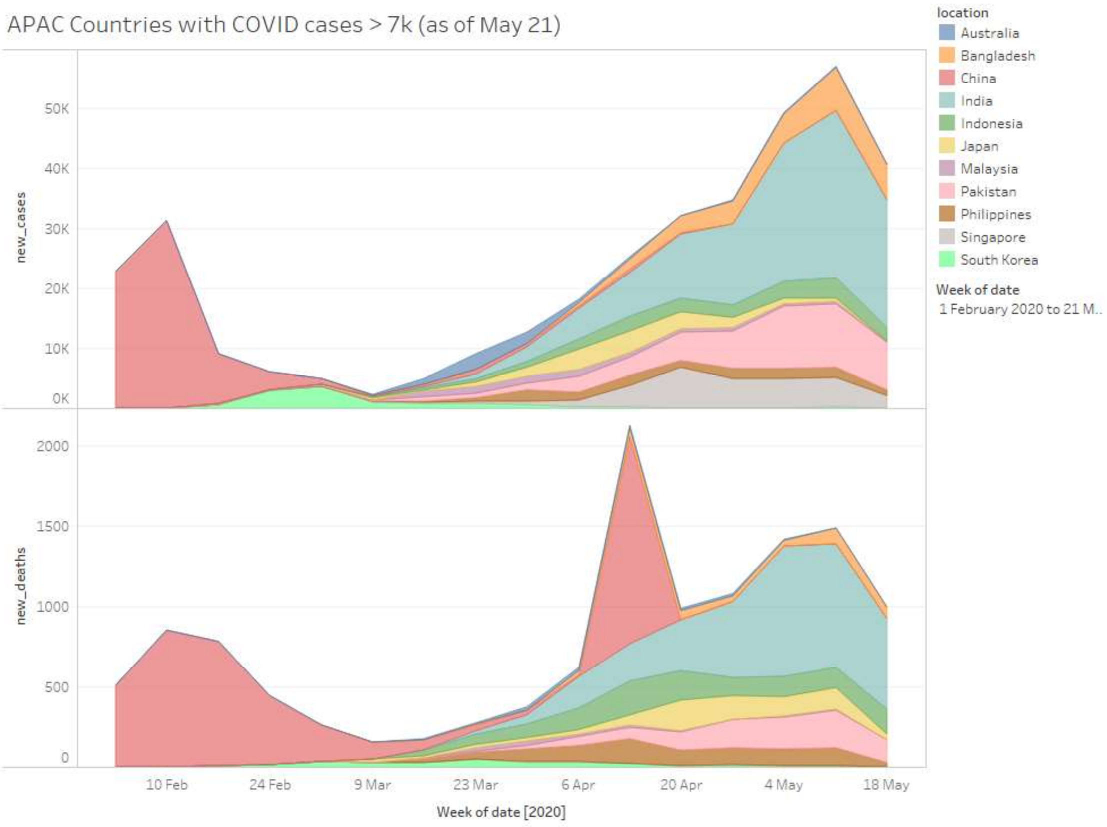
Caption

The size of the pie chart shows the total confirmed cases of CoVID. The red angle on the pie chart pertains to the total no. of death from the confirmed cases.

Figure 8: Total COVID Cases including Death count via pie chart per Country in APAC

To be able to produce that efficiently, a set was configured in Tableau to include all countries under APAC region. In doing this, this does not only provide us a clearer view on the distribution on some countries but also give us another angle to investigate on. To better help us, annotation was used to highlight the attributes of the top 4 countries in APAC with high case of COVID. We see that India has the highest number of cases among APAC countries. Next is China in which it was said to be the virus have originated and have reported 84,067 as of May 21. It is also significant to note that although Singapore may have high cases of COVID but it was able to mitigate the death toll by just having 22 death count.

Using the same set (APAC), another good visualisation to use is the Area Chart. Just like the line graph, this can also show us trend of the case over a period of time. In this case to better improve the visualisation it was filtered to APAC countries with COVID cases higher than 7k. This was also able to address the limitation of Area Chart as having a lot of can create confusion ROM (n.d.).

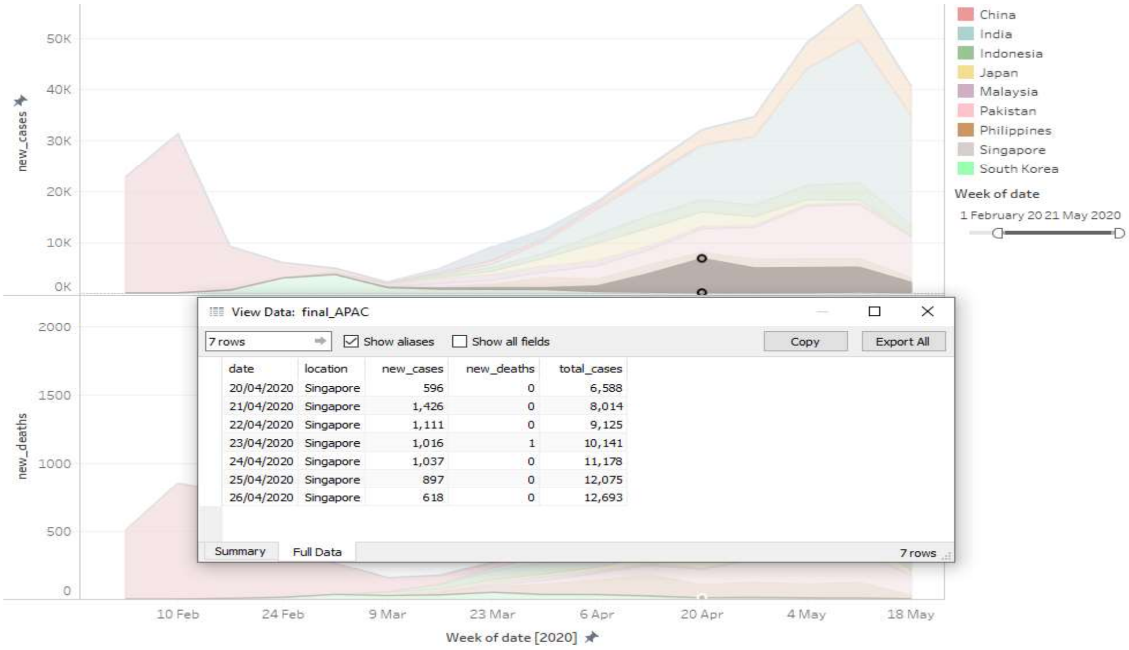


The plots of sum of new_cases and sum of new_deaths for date Week. Color shows details about location. The data is filtered on APAC, which keeps 11 members based on the condition of the total case greater than 7000.

Figure 9: APAC Countries with COVID Cases > 7k with new cases of death using Area Chart

From that figure above, it can be seen that at the onset of the pandemic, China has the greatest number of reported confirmed COVID case and death as well. After March 9, unfortunately, there are countries reporting confirmed cases of the said disease. But before that, in late February, South Korea had already encountered a number of cases in their country as shown by the light green colour in the chart. By late April, it can be observed that India, Pakistan and Singapore have been reporting high number of confirmed COVID cases. On the other hand, in mid-April there was a spike on the number of reported death toll in China. Based on this news ABCNews (2020), it was said that China had to revise the number during that time to include the other cases which were not previously counted, hence the spike. Other countries that can be observed with high death toll on the end of April onwards are India, Pakistan and Indonesia.

In this case, area chart was able to effectively provide more insights regarding the trend of the metrics in APAC regions. Furthermore, from the dashboard, it can still be filtered to focus on particular range of dates. Also, in case the user is curious more on the details of a particular part of the chart, Tableau can also provide additional information by doing a right-click on that region and selecting view data. A sample of this is shown on Figure 10. From the figure below, numbers of each day can be seen. For example, for that week, Singapore's reported new confirmed case on each day is always higher than 500.

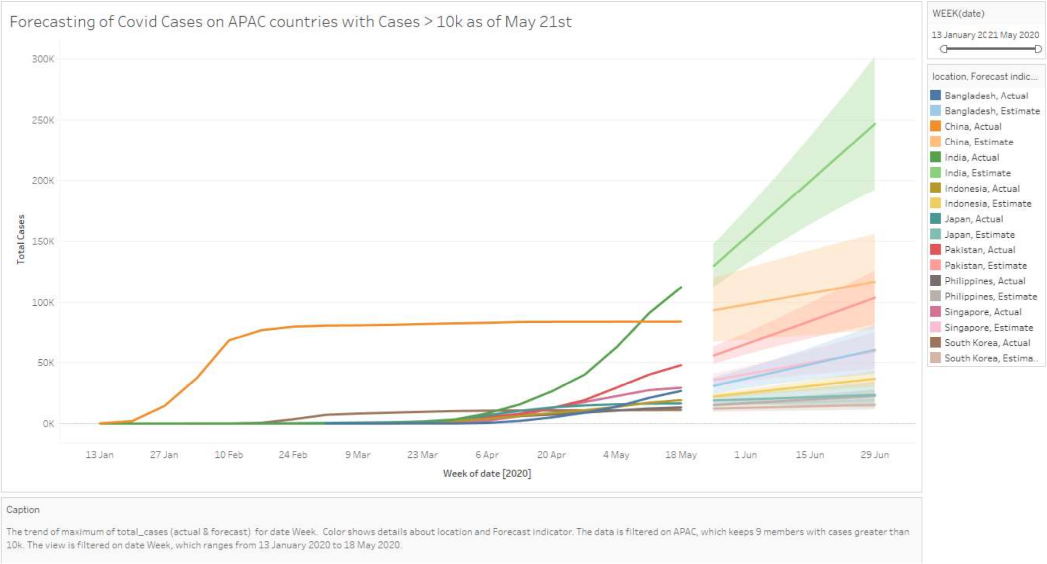


date	location	new_cases	new_deaths	total_cases
20/04/2020	Singapore	596	0	6,588
21/04/2020	Singapore	1,426	0	8,014
22/04/2020	Singapore	1,111	0	9,125
23/04/2020	Singapore	1,016	1	10,141
24/04/2020	Singapore	1,037	0	11,178
25/04/2020	Singapore	897	0	12,075
26/04/2020	Singapore	618	0	12,693

The plots of sum of new_cases and sum of new_deaths for date Week. Color shows details about location. The data is filtered on APAC, which keeps 11 members based on the condition of the total case greater than 7000.

Figure 10: View Data on Singapore COVID Case from Apr 20-26

Another good feature to explore on Tableau is that it has this forecasting technique available for visualisation. In Figure 11, the author used the same set which APAC but had to do filter wherein it just considered countries with confirmed cases greater than 10K. As mentioned in the earlier part, line chart does not work well with too much data. With that set and filter performed forecasting can easily be distinguished among countries. From the same figure, it can be seen that the predicted case would have a steep increase for India, Bangladesh and Pakistan based on the forecasted line. Forecasting is a good proactive approach wherein people will be aware of how to manage things. In this case, government and health officials will be given a guide on how to mitigate the impact of COVID in their respective country.

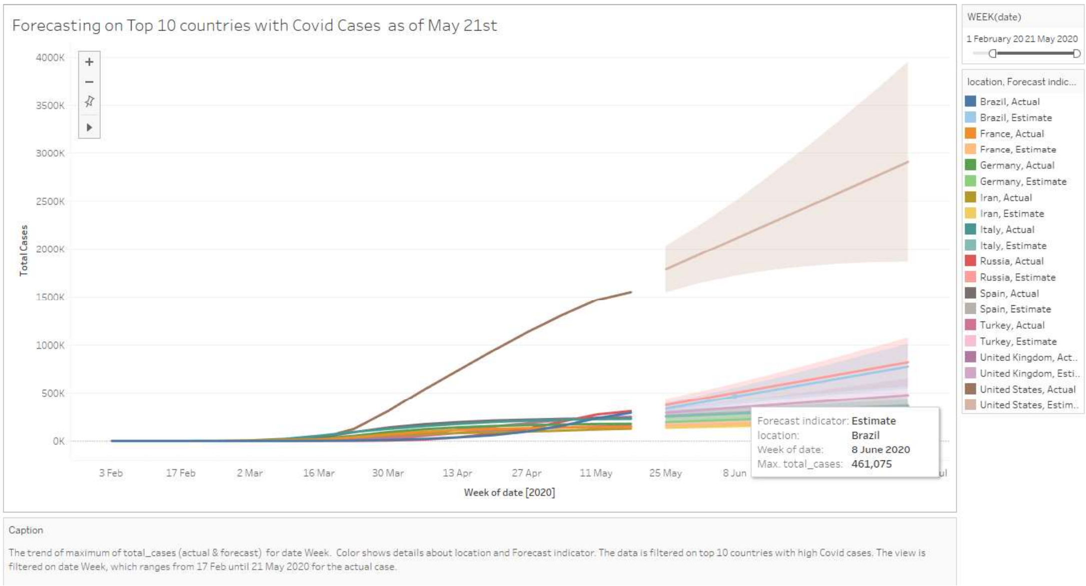


Caption

The trend of maximum of total_cases (actual & forecast) for date Week. Color shows details about location and Forecast indicator. The data is filtered on APAC, which keeps 9 members with cases greater than 10k. The view is filtered on date Week, which ranges from 13 January 2020 to 18 May 2020.

Figure 11: Forecasting COVID Cases in APAC Countries with Cases > 10k

If the user's point of interest is to check the forecasted COVID case for the top 10 countries with high number of confirmed case, this can also be easily be done in Tableau. In this case, the author created another set for countries that belong to this Top 10 category. Similar with the previous approach unique colours were assigned for each country wherein the dark shade refers to the actual count and the light shade corresponds to the estimated count. In Tableau, this is easily done as there is colour palette availble (e.g. Tableau 20) that would automatically do the assignment of colours. In Figure 12, it can be seen on how this was implemented in the said tool. Aside from the obvious that US is forecasted to be have more COVID cases, it also shows that Brazil (blue) and Russia (red) have relatively the same predicted increase of cases.



Forecast indicator: Estimate
location: Brazil
Week of date: 8 June 2020
Max. total_cases: 461,075

Caption

The trend of maximum of total_cases (actual & forecast) for date Week. Color shows details about location and Forecast indicator. The data is filtered on top 10 countries with high Covid cases. The view is filtered on date Week, which ranges from 17 Feb until 21 May 2020 for the actual case.

Figure 12: Forecasting COVID Cases in Top 10 countries

Another information that is good to check on the data set is the test conducted vs total confirmed cases. In this case, another set was created which is to group the eleven countries under Southeast Asia. For the testing however, only the countries shown on Figure 13 have reported pertinent testing information. The other countries were filtered out since comparison will not work. Bar graph can be helpful in looking for insights by checking the percentage of tests with regards to its population and total cases across the seven ASEAN countries. Also, since this just involves small group bar graph should be able to provide a good comparison overview.

It can be seen from there that Singapore is leading with highest percentage of its population being tested at 3.269%. It is also significant to note that the Singapore has the highest number of confirmed COVID case. It may mean that with the ramp in testing that have been conducting they are able to effectively track and detect COVID patients. Also, Malaysia, Thailand and Vietnam have been doing a lot of testing with respect to its population as shown in the graph. It is also significant to note that these three countries having done a lot of testing have lower confirmed total cases as compared to Indonesia and the Philippines. As of this writing, more cases are being reported in Indonesia and the Philippines as they also increase their testing capabilities.

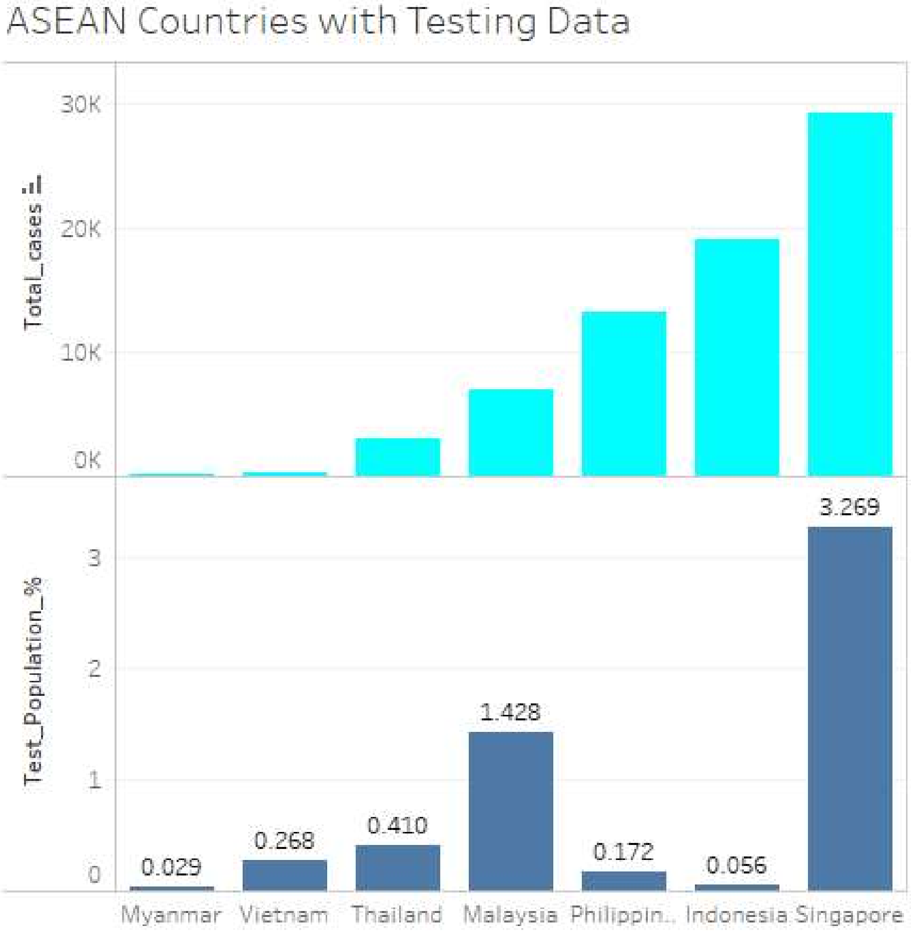
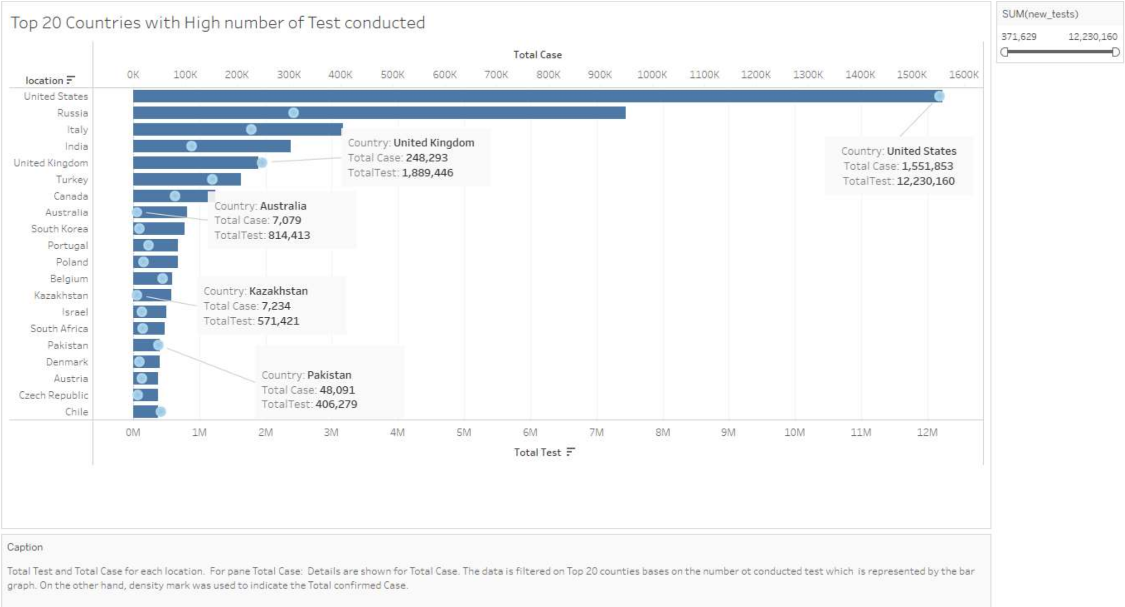


Figure 13: Testing vs Total Cases in ASEAN Countries

The author was also exploring to implement a line chart that could have been better to provide in terms which country is doing an early ramp testing in the previous weeks which could have made an impact in mitigating the impact of COVID. However, there are countries, Singapore for example, without data on the new testing column. Despite that, hopefully, the bar char in Figure 13 is able to provide insights about testing percentage on the population and total COVID cases.

In the next Figure 14, a bar graph was used to show the total number of tests performed but this time integrated with a density marker to indicate the total COVID case. Note that a set was also created to capture the top 20 countries with highest number of tests performed. As can be seen United States which also has the highest number of cases as discussed in previous parts is also leading with the most number of tests conducted which is roughly 12.3 million. Australia and Kazakhstan both having the total confirmed case of less than 8000 are also doing a lot of testing which is higher than other countries having more confirmed cases. Annotation was used in the figure below to show relevant information on the countries mentioned.



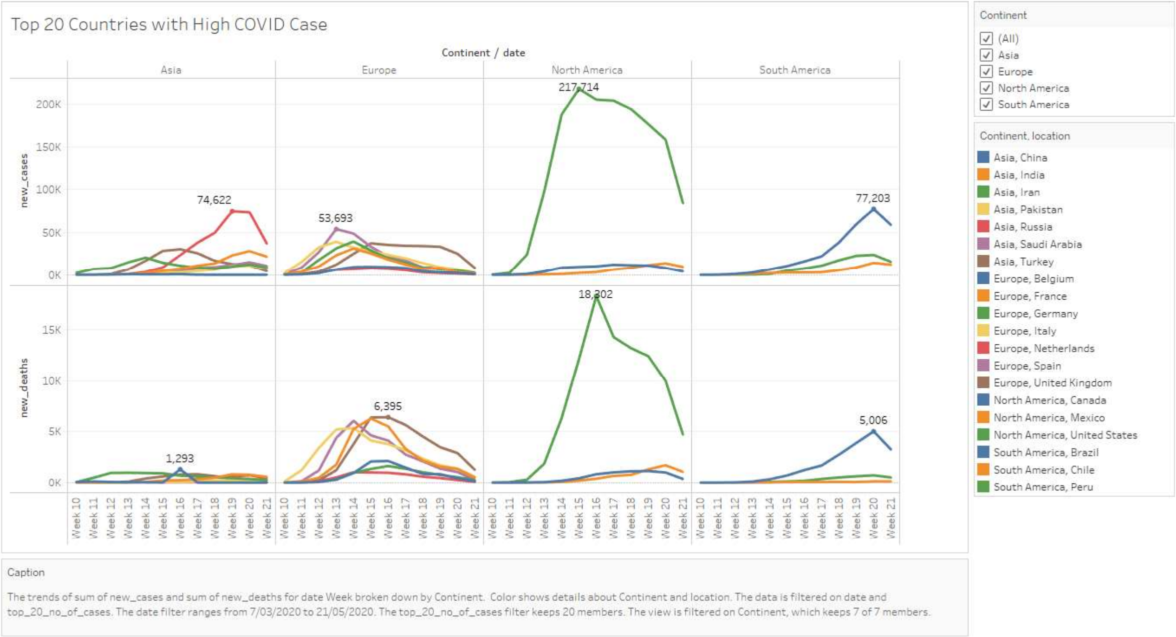
Caption

Total Test and Total Case for each location. For pane Total Case: Details are shown for Total Case. The data is filtered on Top 20 counties bases on the number ot conducted test which is represented by the bar graph. On the other hand, density mark was used to indicate the Total confirmed Case.

Figure 14: Top 20 Countries with High number testing vs Total Case

In Figure 3, line chart was already shown and was also discussed regarding its importance and limitation. It was mentioned there that it is not suitable in presenting big huge amount of data and presenting various dimensions. In Figure 15, instead of presenting them in one graph, each country was plotted according to its continent. In this way it would be easier to analyse the trend per country. Furthermore, it can also provide additional insights by continent. Colour was also uniquely assigned in each county per continent.

Note that Russia for this paper was assigned in Asia. As can be seen on Figure 15, its peak of having high new cases which was recorded on Week 19. In Europe, it can be observed that most of the countries had their highest recorded new cases from Week 12-15. For the European countries shown in the graph, it can be also noticed the drop on the new cases same as on the death toll. However, in South America, particularly for the three countries (Brazil, Peru and Chile), it can be observed that these countries were getting the peak of new cases from Week 18-20. This is also similar to the behaviour of the line for Mexico in North America. Also in terms of death toll, there are more European countries in the set which have high number of reported death due to COVID.



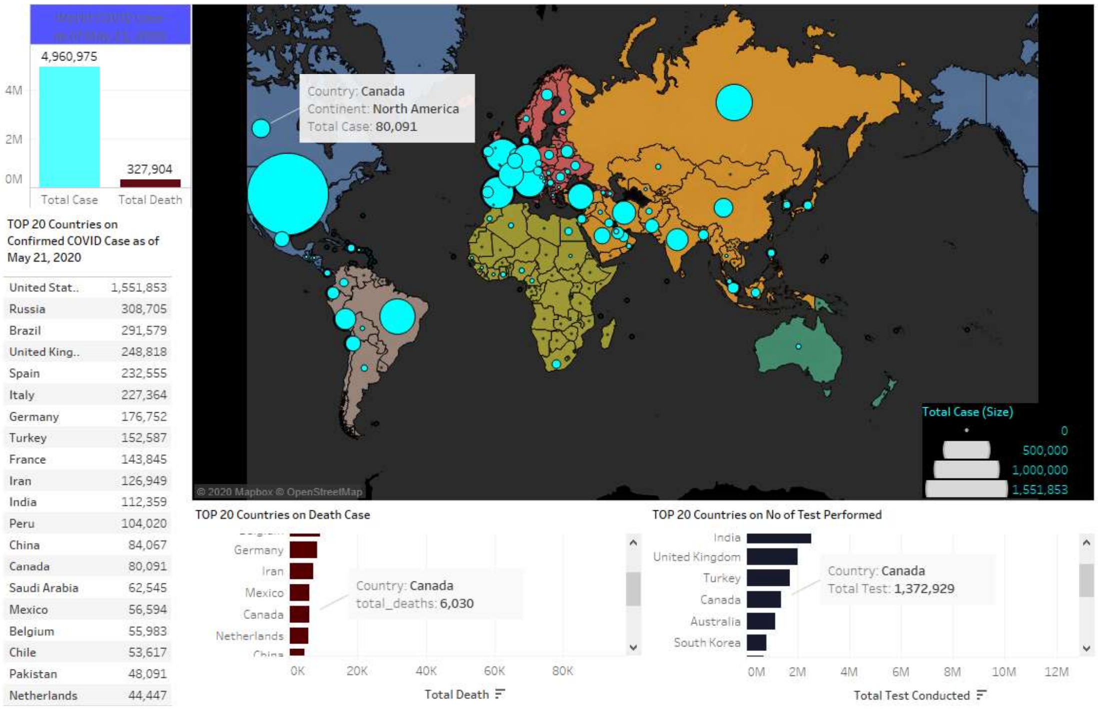
Caption

The trends of sum of new_cases and sum of new_deaths for date Week broken down by Continent. Color shows details about Continent and location. The data is filtered on date and top_20_no_of_cases. The date filter ranges from 7/03/2020 to 21/05/2020. The top_20_no_of_cases filter keeps 20 members. The view is filtered on Continent, which keeps 7 of 7 members.

Figure 15: Line Chart on Top 20 Countries with high case of COVID with new death

For the last part a dashboard, integrating a symbol map, bar chart and table, is implemented. The purpose of having that is to provide an overview of the COVID data set that was provided. In that total cases were shown on the map and table as well. The bar graph was used to show the top 20 countries with highest number of testing conducted and death toll as well. For these two, a scroll bar can be used to check each entry. Note that an annotation was implemented to show that this can be helpful if a user wants to check data of a particular country. In this case, with the annotation, user can easy specific information about Canada's COVID data.

Aside from the graphs shown in this paper, there are other visualisation created which can be seen on the Appendix and some other on the Tableau workbook which was sent together with this paper. For those in the Appendix, those were the initial design but have to be replaced with better visualisation. Like for example, instead of Figure 18, Figure 15 was used since it can better present the line chart by dividing into continents, thus reducing the overlapping of lines.



TOP 20 Countries on Confirmed COVID Case as of May 21, 2020

United Stat..	1,551,853
Russia	308,705
Brazil	291,579
United King..	248,818
Spain	232,555
Italy	227,364
Germany	176,752
Turkey	152,587
France	143,845
Iran	126,949
India	112,359
Peru	104,020
China	84,067
Canada	80,091
Saudi Arabia	62,545
Mexico	56,594
Belgium	55,983
Chile	53,617
Pakistan	48,091
Netherlands	44,447

Figure 16: Dashboard with multiple views of COVID Data

Conclusion

Colour, shapes and labels are among the components in visualisations that were properly considered to ensure the effectiveness of presenting the data in this report. Furthermore, creating set, new calculated fields and additional filtering conditions among others are the additional steps performed to ensure not only to visualize the point of interest but to also to effectively use the graphing techniques such as line, bar and area chart despite its limitation in visualising large amount of data. Providing multiple views using dashboard was also explored in this case it was used to provide an overview of the key metrics in the COVID data set (e.g. confirmed cases, death count and test conducted). Ultimately, the visualisation techniques learned and the guide from Fry (2008) were also applied to effectively provide a meaningful visualisation of the COVID 19 data sets, hence, providing more insights and analysis about the data using graphical representation.

Appendix

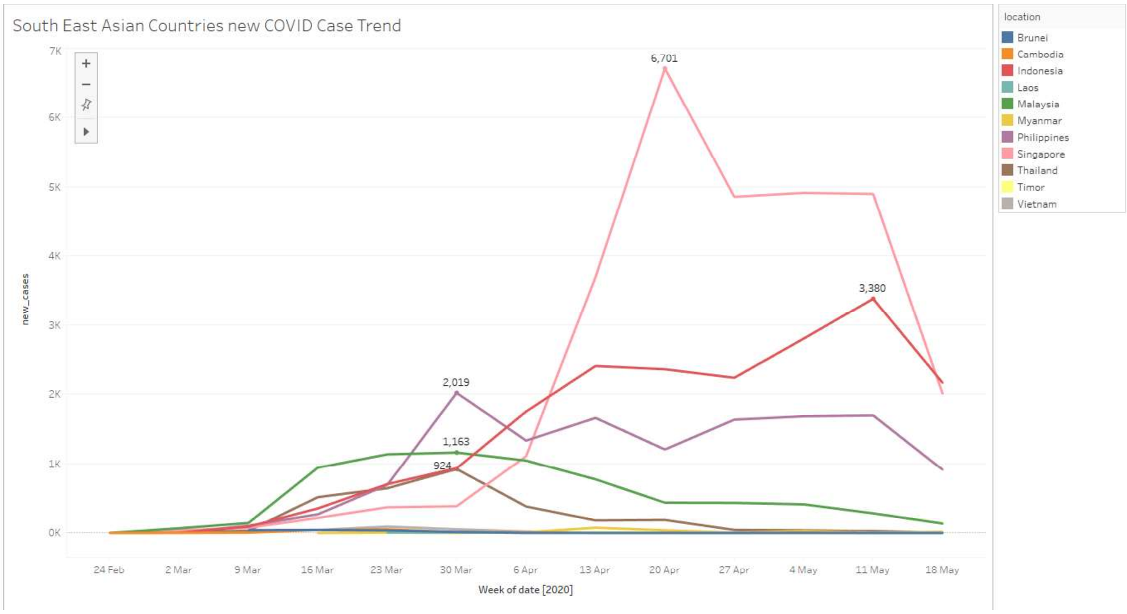
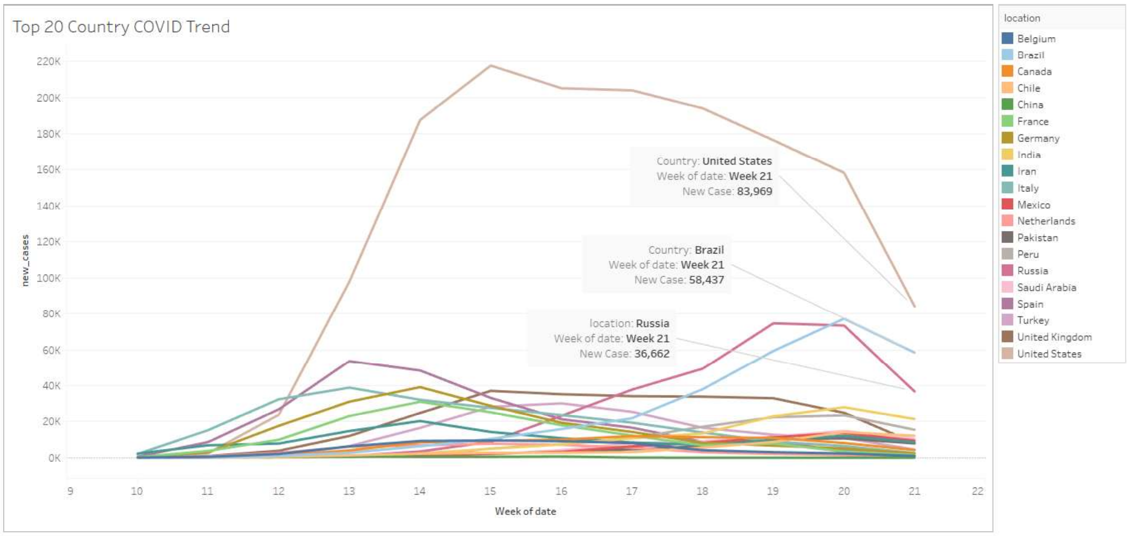


Figure 17: ASEAN New Case Trend



Caption

The trend of sum of new_cases for date Week. Color shows details about location. The data is filtered on date and top_20_no_of_cases. The date filter ranges from 7/03/2020 to 21/05/2020. The top_20_no_of_cases filter keeps 20 members.

Figure 18: Top 20 Country New Case Trend

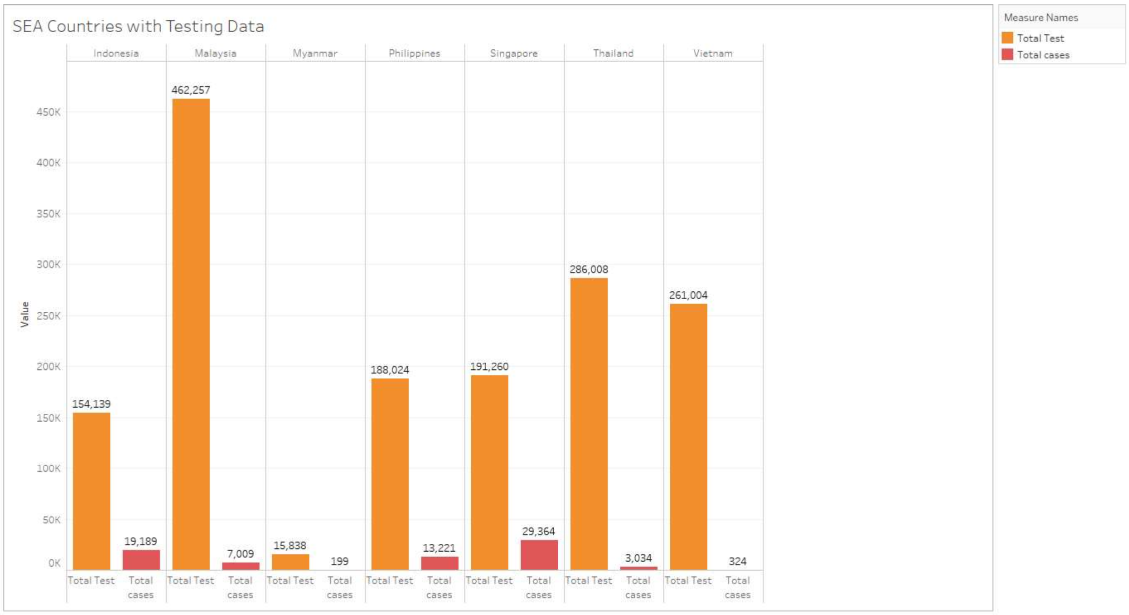


Figure 19: ASEAN Countries with Testing Data vs Total Case

Reference

ABCNews, 2020. ‘China’s Wuhan revises up totial coronavirus death toll by 1,290’. *ABC News*. 17, April 2020, viewed 5 June 2020, <https://www.abc.net.au/news/2020-04-17/wuhan-revises-covid-19-death-toll/12159038>.

Dataflair, 2019. ‘Tableau pros and cons’. *Data Flair*. 18 January 2019, viewed 5 June 2020, <https://data-flair.training/blogs/tableau-pros-and-cons/>.

Fry, B. 2008, July. Data Visualisation. *O’Reilly Media*.

Gulbis, J., 2016. ‘Data visualisation: how to pick the right type’, *eazyBi*, 1 March 2016, viewed 5 June 2020, <https://eazybi.com/blog/data_visualization_and_chart_types/>.

Heitzman, A. 2019, ‘Data visualization: what it is, why it’s important & How to use it for SEO’, *SEJ*, 29 January 2019, viewed 23 April 2020, <https://www.searchenginejournal.com/what-is-data-visualization-why-important-seo/288127/#close>.

Kelleher, C. and Wagener, T., 2011. Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, *26*(6), pp.822-827.

Kiss, M., 2011. ’10 Tips for Presenting Data’. 18 January 2018, viewed 5 June 2020, <https://resources.observepoint.com/blog/10-tips-for-presenting-data>.

Mace, C., 2014. ‘Show me how: symbol maps’. *Information Lab*, 11 December 2014, viewed 5 June 2020, < https://www.theinformationlab.co.uk/2014/12/11/show-symbol-maps/ >.

Plaisant, C., Schneiderman, B., Chintalapani, G. and Aris A., 2004. Treemap Project. HCIL:University of Maryland, viewed 23 April 2020, <http://www.cs.umd.edu/hcil/treemap/>.

Ritchie, H., 2020. ‘Coronavirus source data’. *Our World in Data*. May 2020, viewed 21 May 2020, <https://ourworldindata.org/coronavirus-source-data>.

ROM, n.d. ‘Advantages and disadvantages of different types of graphs’ *Rom Knowledgeware*. Viewed 5 June 2020, < http://www.kmrom.com/Site-En/Articles/ViewArticle.aspx?ArticleID=416 >.

Senay, H. and Ignatius, E., 1994. A knowledge-based system for visualization design. *IEEE Computer Graphics and Applications*, *14*(6), pp.36-47.

Shaffer, J., 2017. Using treemaps to visualize data. *Data Plus Science*. 3 April 2017, viewed 23 April 2020, <https://www.dataplusscience.com/UsingTreemaps.html>.

Shneiderman, B., 1992. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics (TOG)*, *11*(1), pp.92-99.

Stasko, J., Catrambone, R., Guzdial, M. and McDonald, K., 2000. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International journal of human-computer studies*, *53*(5), pp.663-694.

Tableau n.d., ‘What is Tableau’. Viewed 5 June 2020, <https://www.tableau.com/products/what-is-tableau>.