**Visualization of COVID-19 data set with Tableau**

**Introduction**

Visualization is defined as the use of computer-aided, interactive, visual representations of abstract data to increase cognition (Card et al. 1999) Data visualisation is a graphical analysis of the quantitative details. In other words, data visualisations transform big and tiny data collections into images which are easier to interpret and process for the human brain. Visualizations of the data may be used to track hidden information and patterns. The goal of visualisation techniques is to represent the multidimensional data in a low-dimensional space in order to maintain properties like clusters, outliers of the data set structure.

The focus of this paper is to identify and develop visualisations for covid-19 multidimensional dataset and get a clear picture of growing coronavirus spread. It will also focus on exploring the dataset to find insight, trends and patterns from the visualisations.

Data visualisation is important, since it makes it easier to see trends and patterns. This helps decision-makers to act more quickly based on the data visualized and presented. Analysis can be made simpler with interactive visualisation, by using visualizations to interpret charts and graphs for more detail, interactively manipulating what data you see and how it is processed.

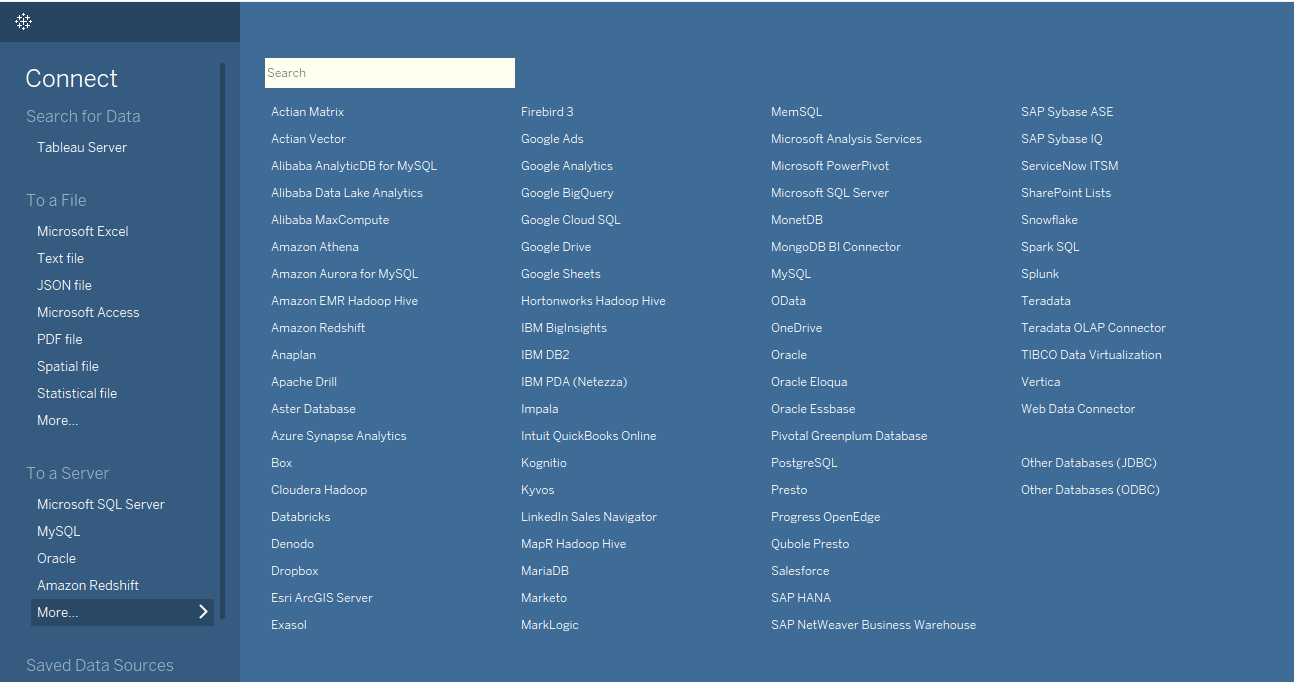
**Visualization Tool – Tableau**

Tableau is a visual analytics platform transforming the way we use data to solve problems and empowering people and organizations to make the most of their data (Tableau 2020). It is a software tool which is used to visualize multidimensional datasets. It is a business intelligence software which makes data visualization and analytics easy with drag and drop features. There is no prior programming experience required to use Tableau.

Tableau removes needless complexities, focussing on what matters most that is to find meaning in the data. It offers a range of visualisation tools to make the users' data more available. Tableau is a new age data analytics and visualizations tool that provides flexibility and ease-of-use with a smooth experience to the users (Nikhat et al. 2020).

**Data Sources:**

Tableau gives multiple options to connect to your data as shown. It provides flexibility to connect to multiple data sources at a time, supporting big data eco-system, integration with R for statistical analysis, connection with Non-RDBMS data sources like JSON and web data connectors to connect to different sources helps Tableau to stand out.



*Figure 1 Tableau Data Sources*

**Multidimensional Data**

Multidimensional analyses include identifying similarities between data dimensions, studying data clusters and identifying outliers (Gregorio 2016). Its data classes are divided into two categories: data and measurements (Wikipedia, Multivariate Analysis 2016).

Therefore, we have multidimensional data and our goal is to make some visual insight into the data set analysed. For human perception, the data must be represented in a low-dimensional space, usually of two or three dimensions (Dzemyda & Kurasova 2015).

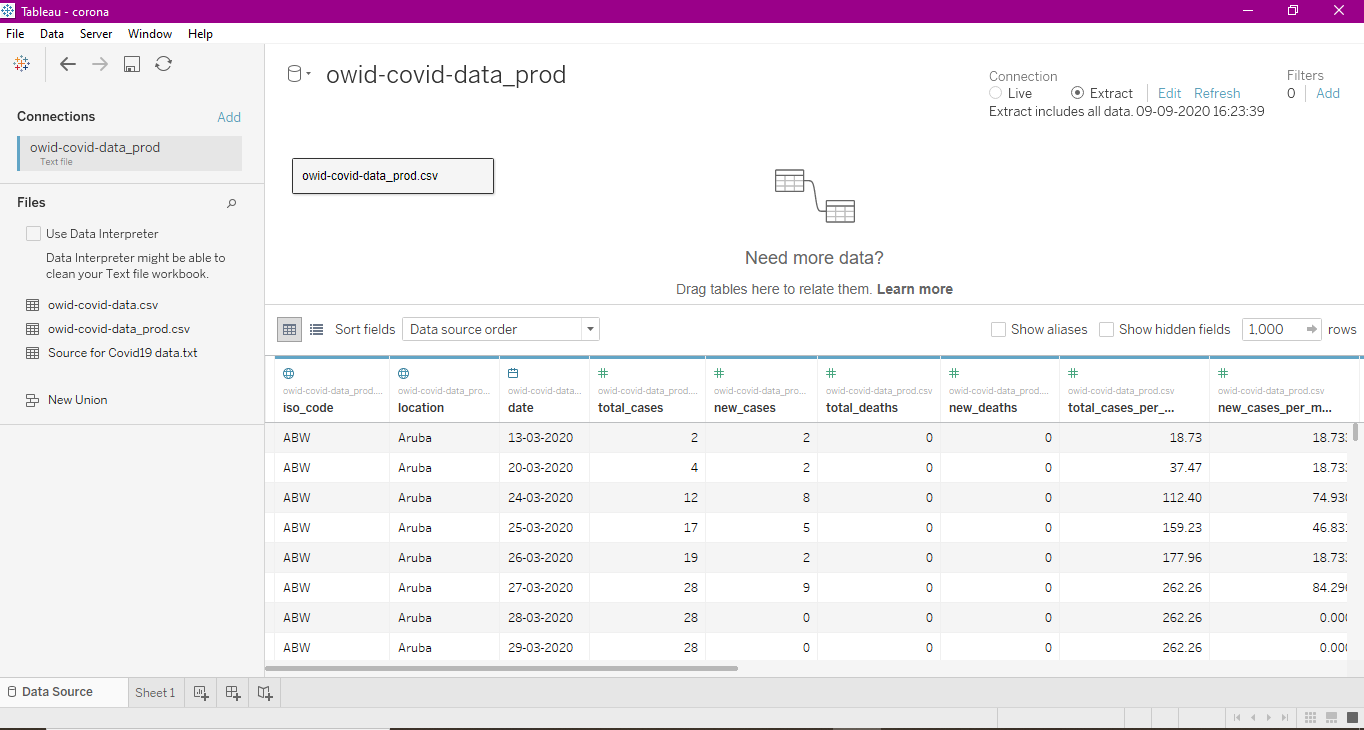
For this paper the following dataset is used: -

**Coronavirus Data Set**

The coronavirus dataset is a multidimensional dataset that contains information of covid-19 cases in all countries around the world from December 2019 to April 2020.The data is from the European Centre for Disease Prevention and Control (ECDC). The dataset has 13467 entries and the 16 columns.

**Loading the Data**

The owid-covid-data.csv was loaded into tableau using the following steps: Connect->To a File-> Select the csv. Tableau automatically detected first row as column names and separated quantitative columns as **Measures** and categorical columns as **Dimensions.**



*Figure 2 Loading data into Tableau*

During visualization of data as a geographical graph, a geographical role was given to the dimensional values “iso\_code” and “location”. The country “geographical role” denoted with a tiny globe icon in Tableau, the centre points of each geographic unit will automatically produce longitude and latitude values. It was found that Tableau was not able to map certain locations as they were being referred as different name in Tableau for which Edit Location dialog box was opened to enter corresponding country name and a filter was applied for “world” and “international” value.

|  |  |
| --- | --- |
|  |  |
|  |  |

*Figure 3 Mapping location data in Tableau*

**Visualisation and Analysis**

**Geospatial/Geographic Visualisation**

Geographic visualisation involves geospatial data processing methods and techniques using immersive visualisation. Geo visualization integrates approaches from visualization in scientific computing, cartography, image analysis, information visualization, exploratory data analysis and geographic information systems to provide theory, methods, and tools for visual exploration, analysis, synthesis, and presentation of geospatial data (MacEachren & Kraak 2001).

It will be important to analyse geographic visualizations to study the impact on different parts of world from coronavirus pandemic.

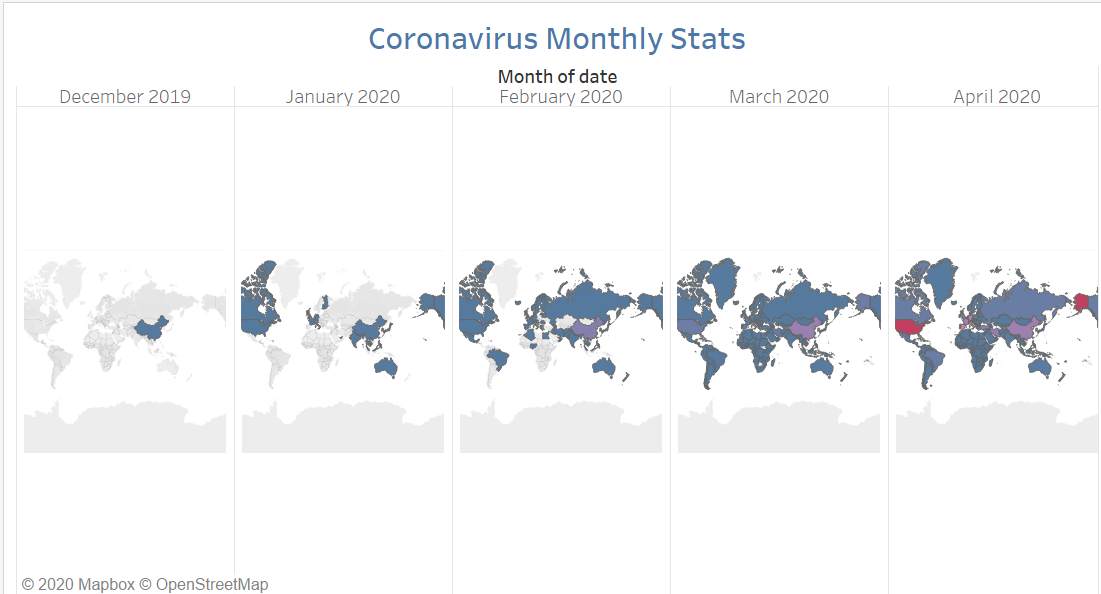
**Choropleth Maps**

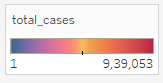
A choropleth map uses variations in filtering, painting, or putting objects inside predefined areas to display the average property or quantity values in those regions. (Lexico 2020). Choropleth maps usually uses different hues of one colour or gradient of one colour to another.

**Advantages: -** They are very easy to understand and show spatial distributions of data quite well. They help to identify interesting hot points, to detect associations between the encoded component and geographic location.

**Disadvantages: -** The data values are not necessarily correlated with the areas. They offer the false appearance of sudden shift at shaded unit boundaries. Most choropleths are not fit to display absolute values.

**Analysis:**



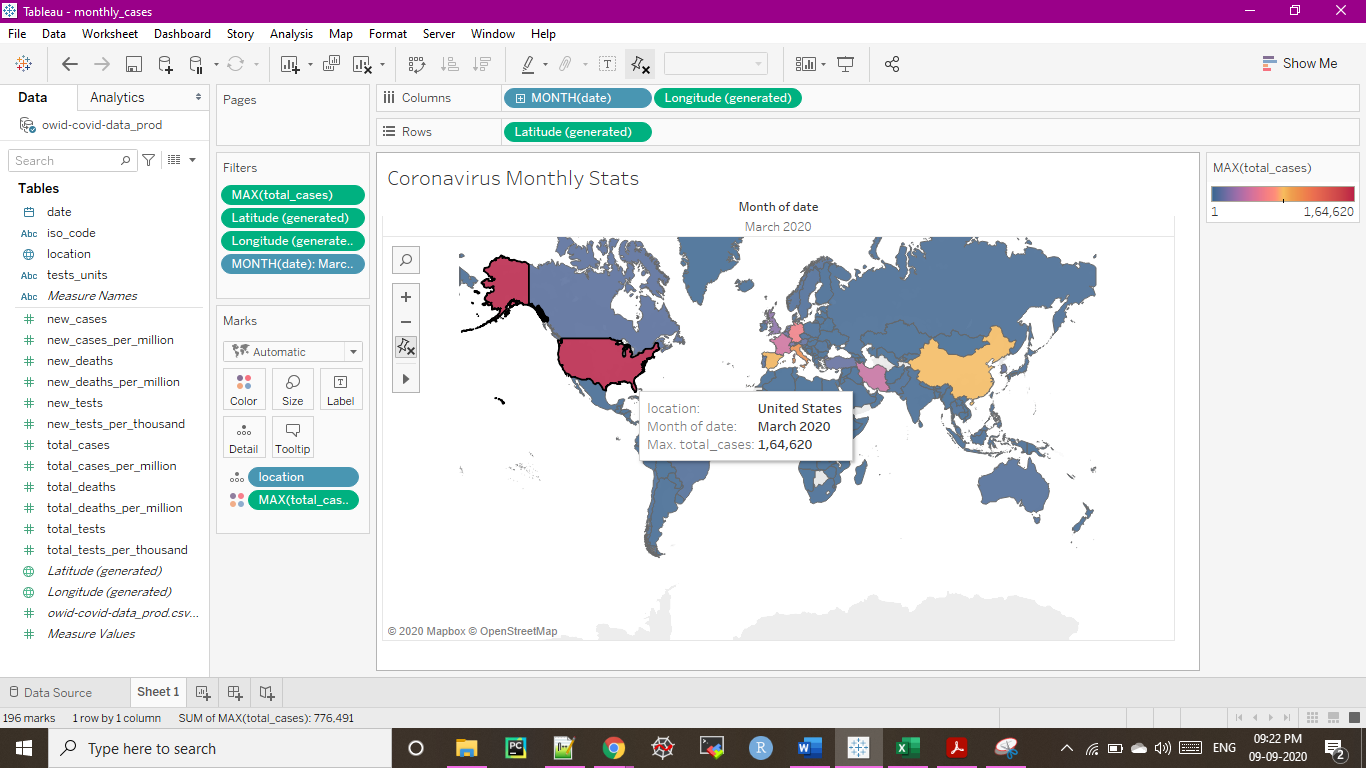


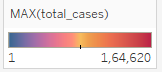
*Figure 4 Choropleth Map Covid-19 month wise*

The visualization shows that the **spread of coronavirus in world started from China** in the month of December with 27 cases, which gradually increased in February due to international travel of people from one country to another and through community spread captured almost all the countries in the month of March and April 2020.

It is also observed that USA, Canada, India, Australia, UK, France were some of the countries where the spread began first during the month of January 2020.In March **USA with 164620 cases surpassed China (82241) to become the country most effected by coronavirus** in the world.

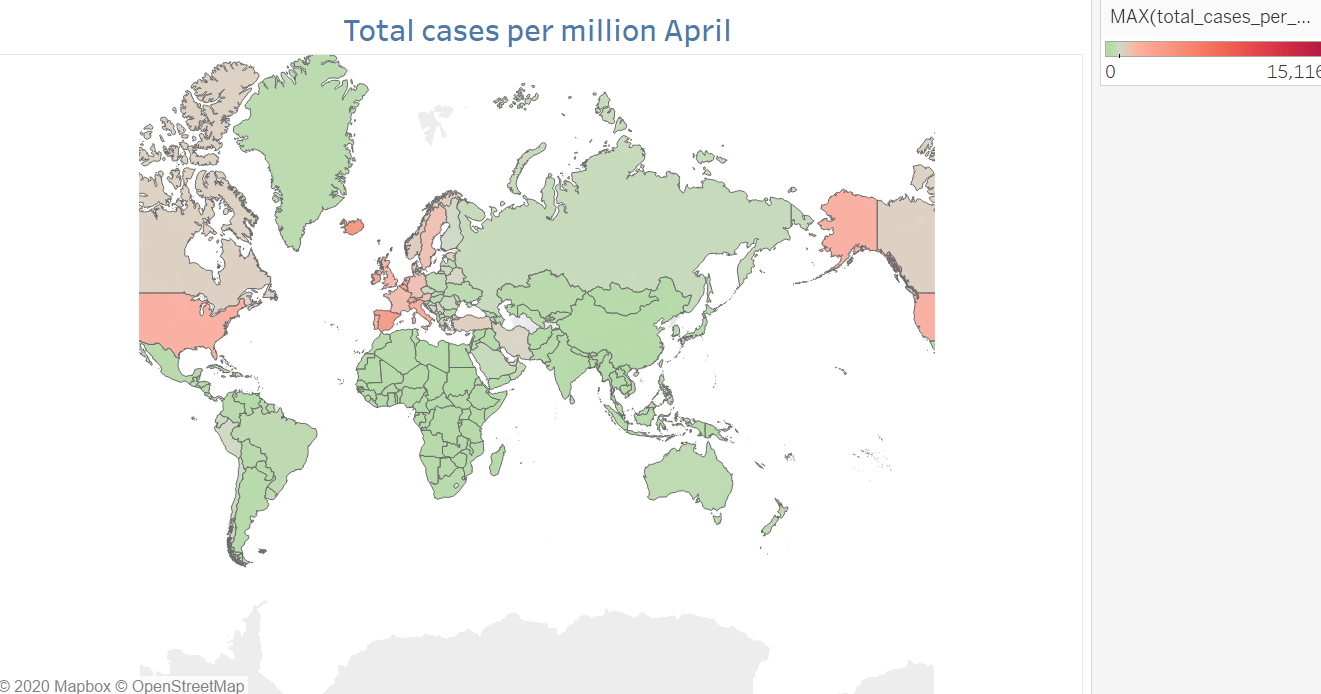
The Blue colour indicates the counties with less cases, purple with high and red with highest cases.





*Figure 5 Covid-19 cases in USA in March 2020*

Italy and Spain also become heavily infected by coronavirus with 101739 and 85195 cases respectively. In April coronavirus continued to spread exponentially throughout the world.



*Figure 6 total cases per million in April*

We cannot estimate the measure of spread with no. of cases alone because of different populations of different countries so it’s better to analyse cases per million to have a better estimate on situation of each country. The above figure shows that USA, Spain, Italy, Iceland has the highest cases per million of the population in month of April. It shows that these were the countries which were struggling the highest from the pandemic. The green colour represents the countries least affected and red with most affected.

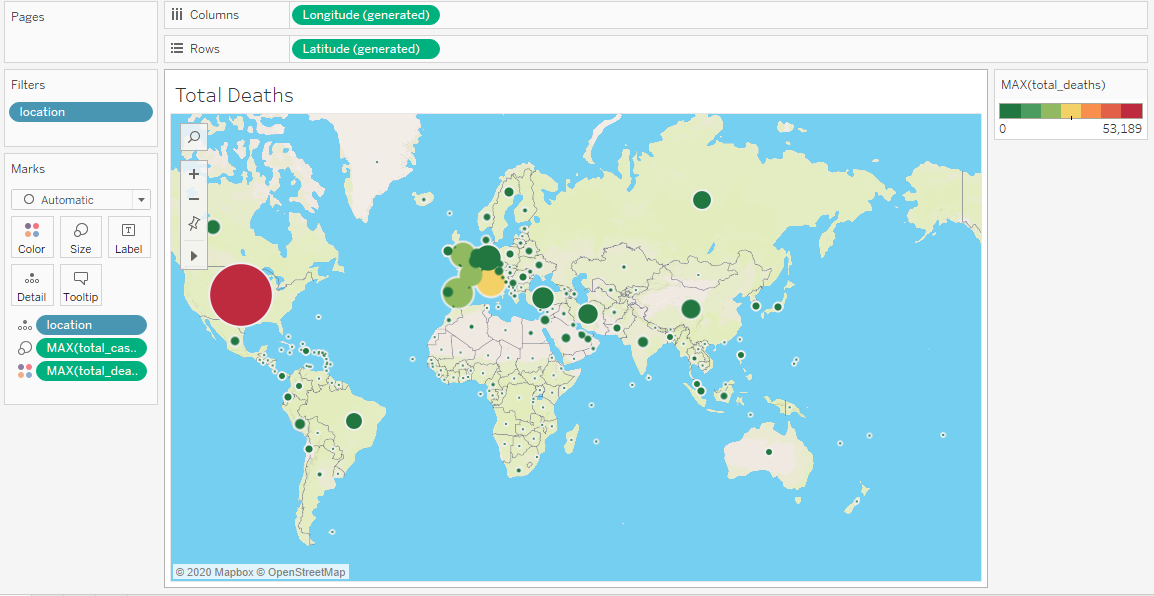
**Symbol Maps**

Symbol Maps use the visual size variable to reflect variations in the magnitude of a subtle, suddenly shifting phenomenon. Proportional Symbol Maps are unclassified variant of maps, where the symbol area is proportional to the mapped attribute values. Proportional maps of symbols scale the size of simple symbols proportional to the value of data contained at that spot. The larger the symbol at a spot, the large the value. The simplest approach is to directly scale the circles proportionate to the results.

**Advantages: -** Proportional maps of symbols are very versatile, since either numerical data or categorical data ordered can be used. They are also versatile, since data attached to geographic points or data attached to geographic regions may be used.

**Disadvantages: -** A common problem with proportional symbol maps is symbol congestion / overlap, especially when there are wide differences in symbol size or when there are multiple data locations close together. Another common issue with proportional symbol maps is that map readers usually don't very well estimate the symbol areas.

**Analysis:**



*Figure 7 Symbol map total no. of cases vs total no. of deaths*

The size of bubble represents the total no. of cases and the colour of bubble depicts the no. of deaths. We can observe that the bubble of Italy is smaller than that of Spain, but the colour depicts more deaths. Also, the average age of Italy is higher than average age of Spain (Worlddata 2020). The bubble size of which **suggests that older people have high risk of death** but needs further investigation. USA has most no. of cases and most no. of deaths than any other country.

|  |  |
| --- | --- |
| **Average Age of Italy 2020** | **Average Age of Spain 2020** |
| 45.4 years | 41.4 years |

**Heat Maps**

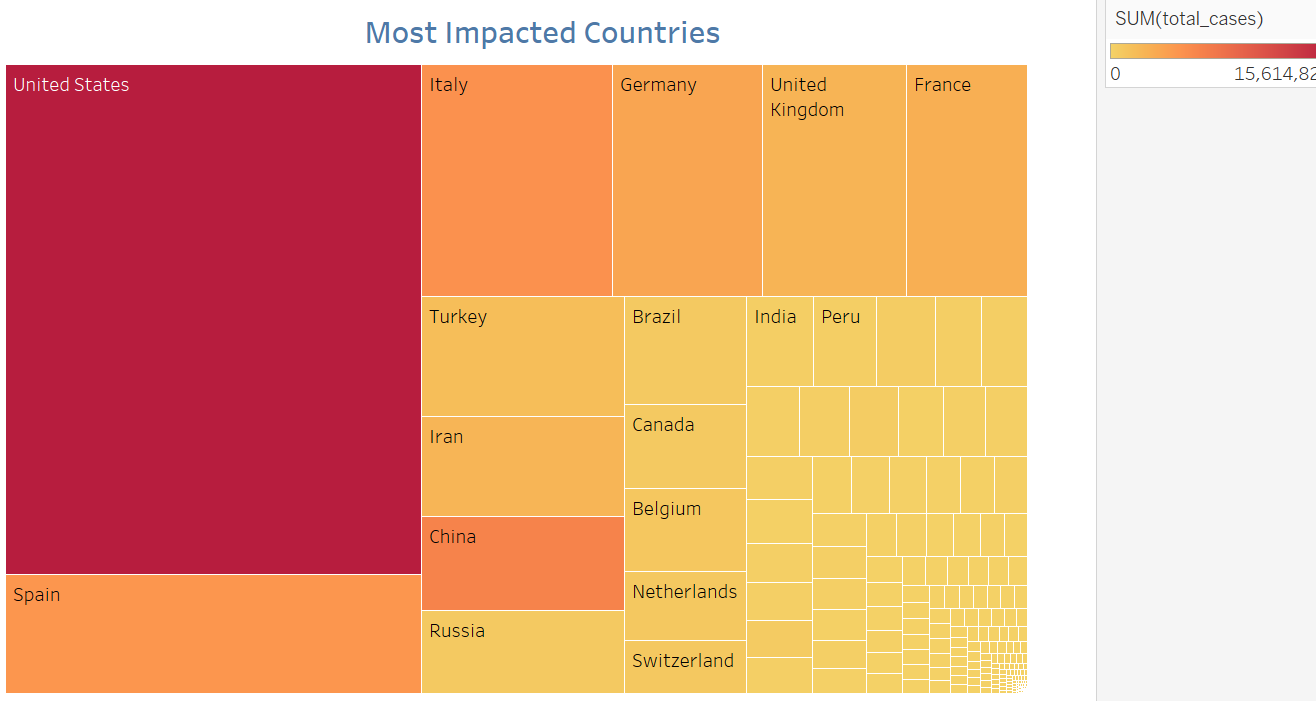
A heat map is a graphical representation of data where the individual values contained in

a matrix is represented as colours (Moon 2016). It allows you to view some variable (in colour) as a function of two other variables on the x and y axis.

**Advantages: -** The use of colour gradients and consistent structure make it easy for viewers to interpret the chart on different scales, from the individual data point to the whole data.

**Disadvantages: -** As the number of items increase in a tree map, the amount of space allocated for each item decreases.

**Analysis:**



*Figure 27 most impacted countries*

The Heat/Tree map clearly shows the most impacted countries due to the pandemic with most cases of coronavirus – United States, Italy, Spain, Germany, United Kingdom, France.

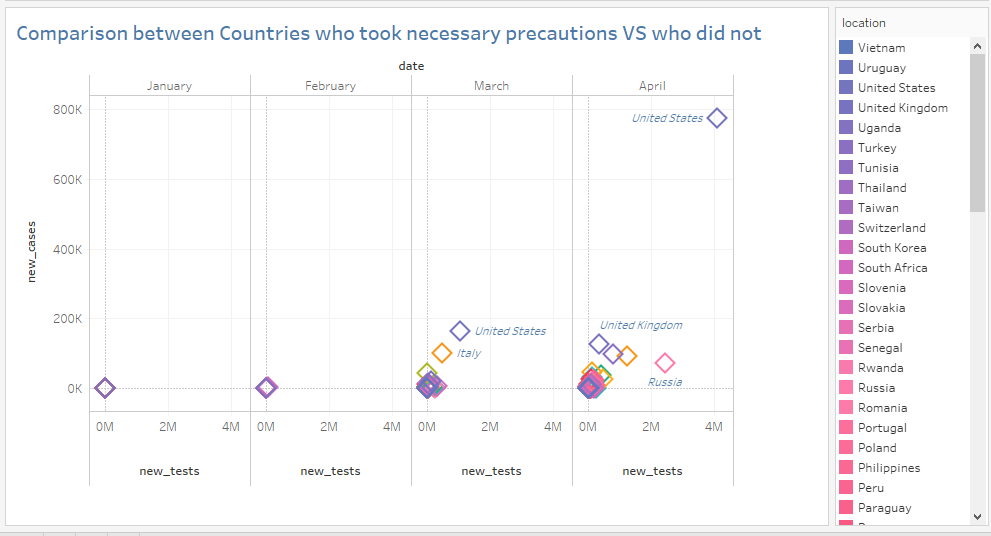
**Scatter Plots**

Scatter Plots are one of the most powerful and most widely used techniques for visual data exploration. (Keim et al. 2010). Scatter plots are like line graphs in that they map data points using horizontal and vertical axes. Scatter plots demonstrate how influenced one variable is by another. The association between two variables is called a correlation between them.

**Advantages: -** It is the best way to visualize a non-linear pattern, and you can evaluate the range of data flow, i.e. maximum and minimum value.

**Disadvantages: -** They can't provide the exact extent of the correlation.

**Analysis:**



*Figure 28 Scatter plot of each month*

During the month of March, we observed that countries such as Italy, Australia, India who took some serious **precautions like strict lockdowns and which implemented social distancing guidelines in the mid of March has less new incoming cases in April but in some countries such as USA were strict guidelines were not implemented during month of March**, experienced very high jump in new cases during April.

Also, the number of new tests in April is increased all over the world but the jump is higher in countries under lockdown.

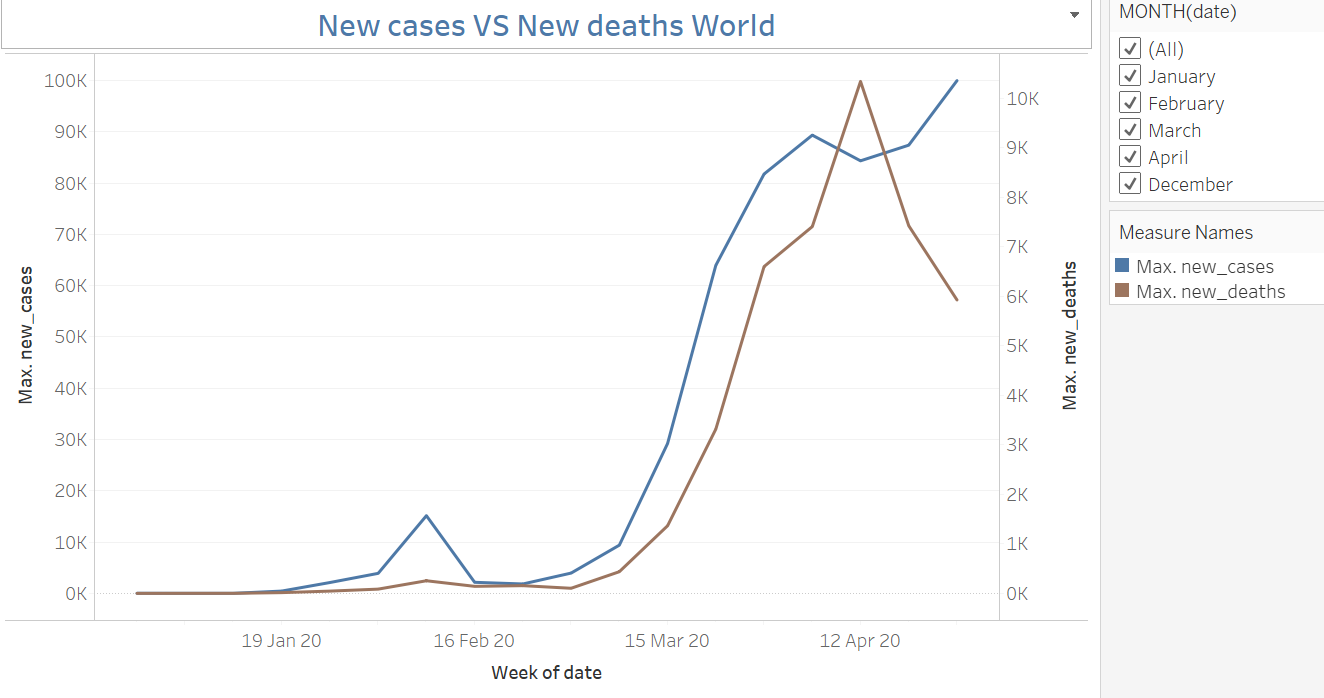
**Line Charts**

A line chart is a type of chart which displays information as a series of data points called markers connected by straight line segments (Wikipedia, Line charts 2020).

**Advantages:** - They provide simple presentation and are easy to read and create. They are good for showing trends over periods of time. They can also handle both negative and positive values.

**Disadvantages: -** They do not work well for categorical data and are harder to read when lines overlap frequently.

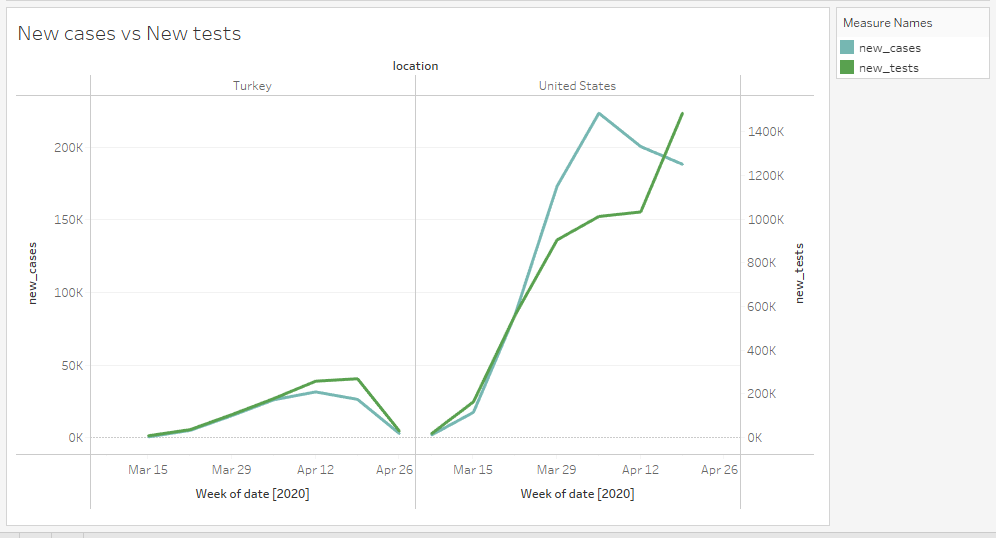
**Analysis:**



*Figure 8 new cases vs new deaths*

The above visualization shows new cases confirmed per week with new deaths recorded per week all other the world. Maximum is used as the aggregation function hence; each data point shows the highest new confirmed cases reported each week and highest new deaths reported each week.

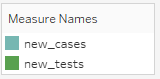
It is observed that the new cases reported each week has been growing linearly but the new deaths reported each week doesn’t have much slope hence is not as drastic as new cases and are coming down since April 12th, 2020.



*Figure 9 new cases vs new deaths for Turkey and United States*

A few countries such as United states claimed that they reported a higher no. of cases because more tests are being conducted which was shown in visualization of sum of new cases each week vs sum of new tests each week. From graphs of Turkey and United states we can observe that the new cases have increase as new tests are being conducted.

|  |  |
| --- | --- |
|  |  |

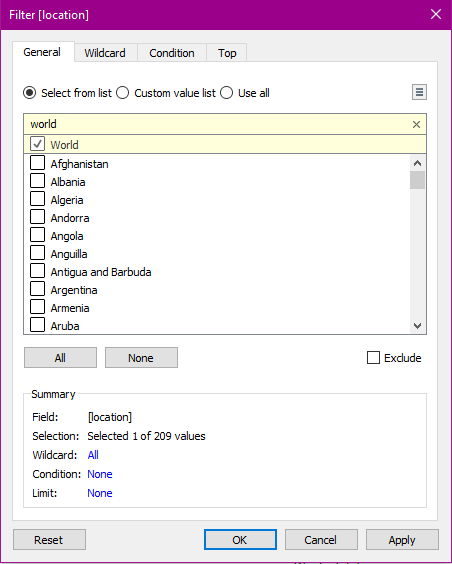


*Figure 10 new cases vs new deaths for Italy and New Zealand*

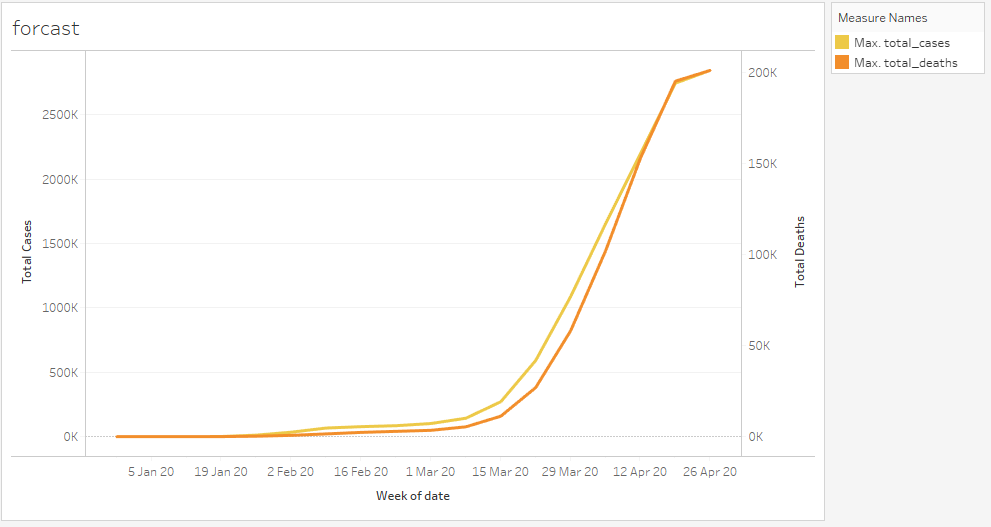
But from visualizations of Italy and New Zealand we can observe that there **is no correlation between sum of new tests performed each week and sum of new cases each week**. During certain weeks the number of new tests has increased while there is a decrease in new cases. This claim was further investigated by the author with help of bar plot visualizations.

**Forecast**

Below graph shows total deaths, total cases, total new cases throughout Dec – April from all over the world by filtering location to only include data of world.

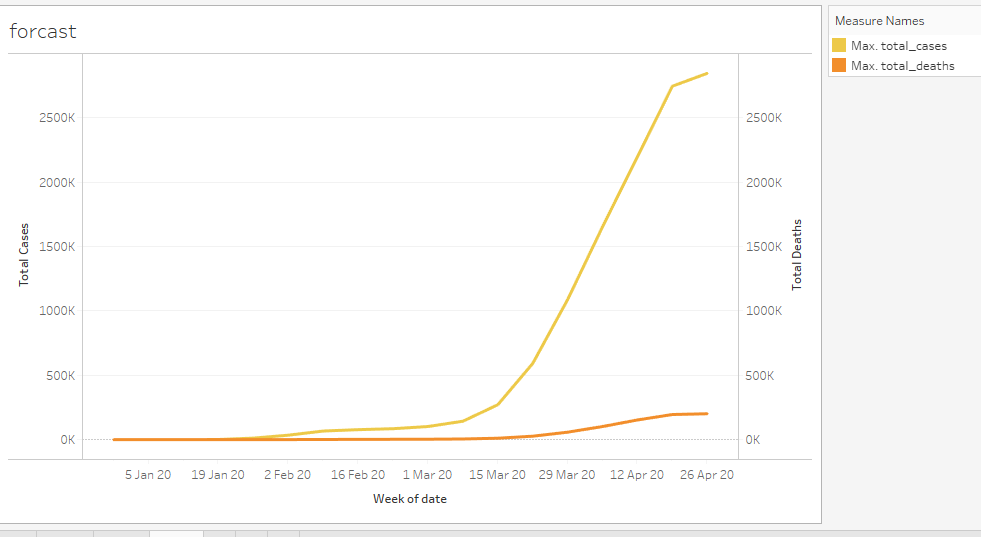


*Figure 11 filter Location world*



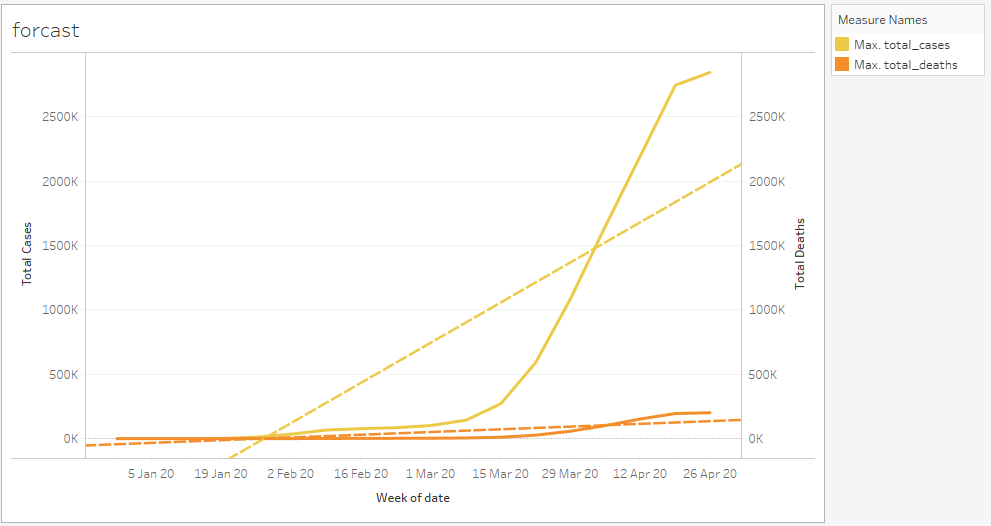
*Figure 12 Total Cases vs Total deaths for world*

As we can observe there is a liner relationship between both measures but to make more sense of data both of the axis were synchronized with each other for proper interpretation.



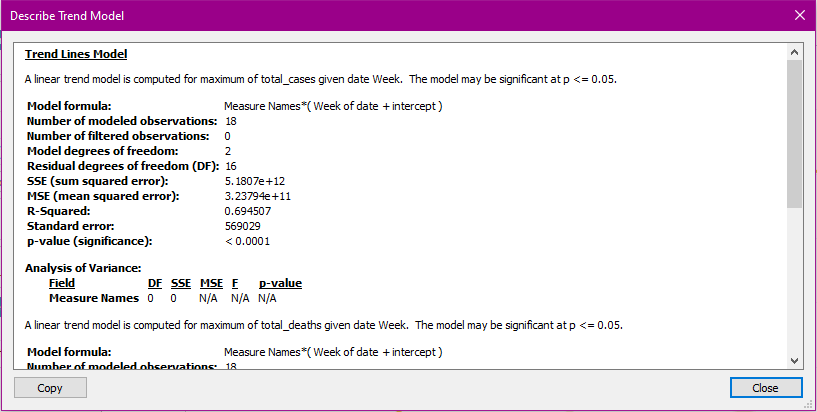
*Figure 13 Total Cases vs Total deaths for world synchronized axis*

We can observe that total deaths are increasing rapidly but total cases of coronavirus are increasing exponentially in the month of April. Till February there was a steady growth but from the month of March the cases have increased dramatically.

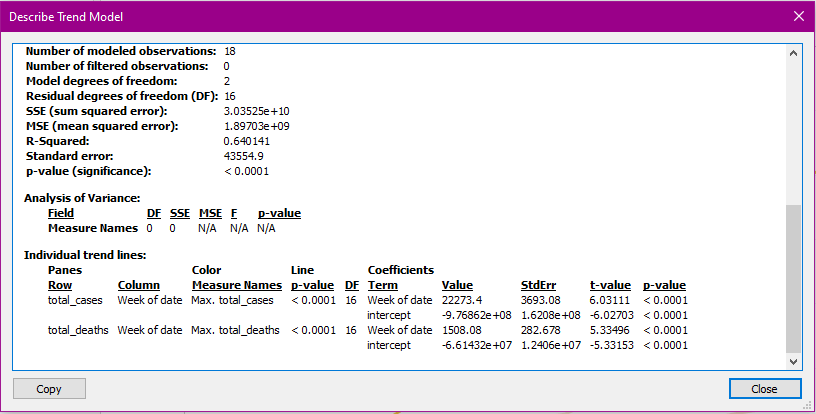


*Figure 14 Total Cases vs Total deaths Trend lines*

In Tableau, we get a lot of options to create different types of trend lines which helps in inferring important patterns and trends in our data. Trend lines allow you to interpret the data by placing a straight or curved line that best represents the pattern found in detailed data plots. They also enable you to see data patterns that are not visible when viewing the source data chart by drawing a line that best fits the values in view. Trend line were added to the chart with prediction confidence intervals of 95%. The trend model details are: -



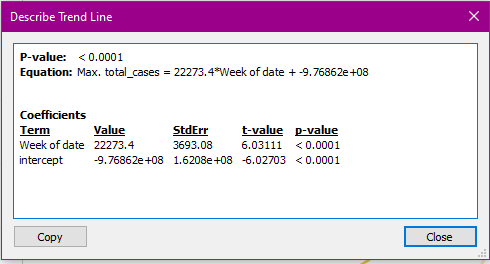
*Figure 15 Trend model details 1*

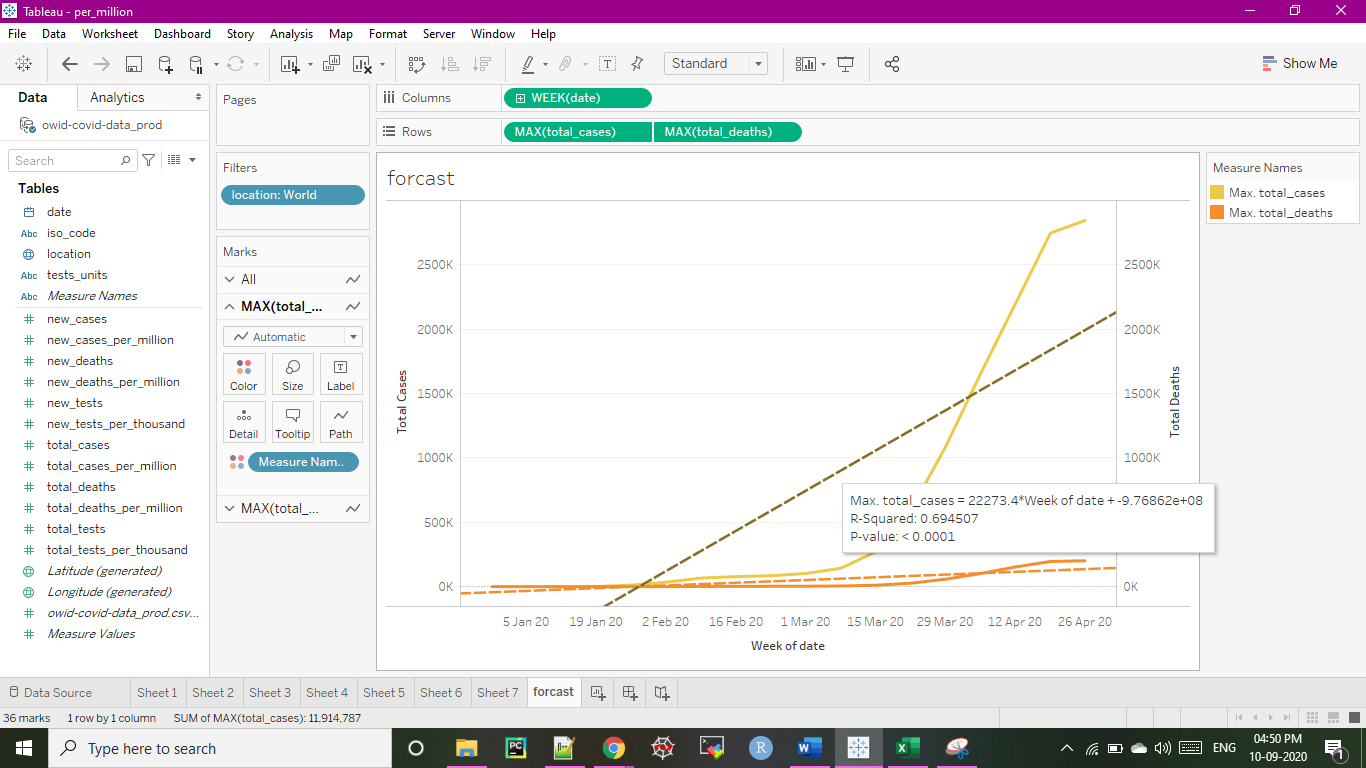


*Figure 16 Trend model details 2*

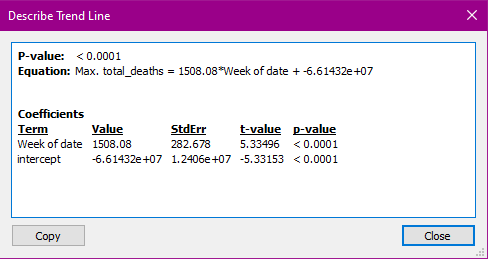
From trend lines we can observe that there is a linear increasing trend in both total deaths and total cases. The linear equation for this model is: -

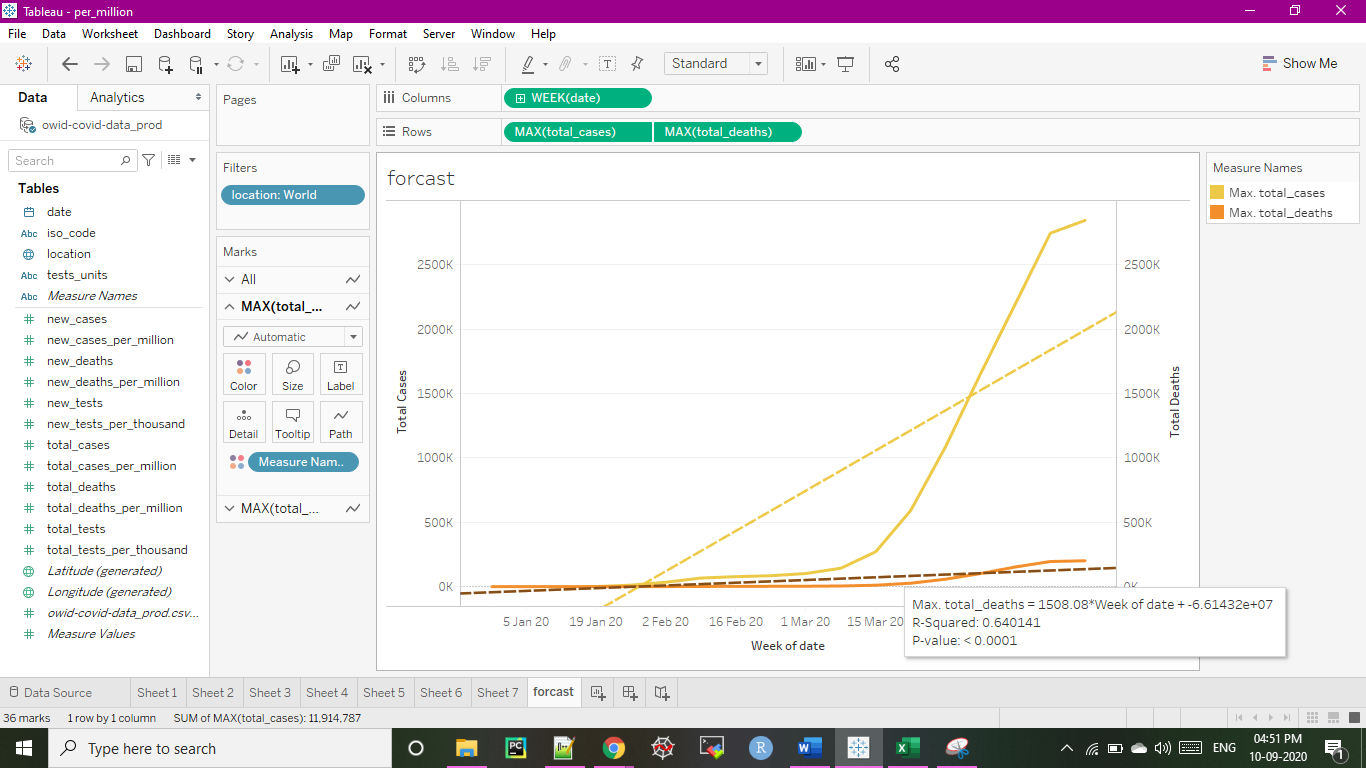
**Measure Names\*Forecast indicator\*(Week of date + intercept)**





*Figure 17 Trend line description for total no. of cases*



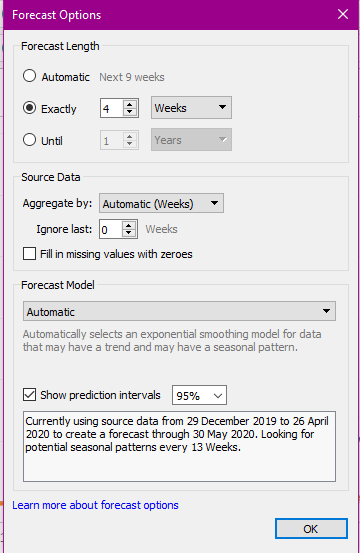


*Figure 18 Trend line description for total no. of deaths*

The p‐value for the regression is less than 0.001, which shows that the model has good

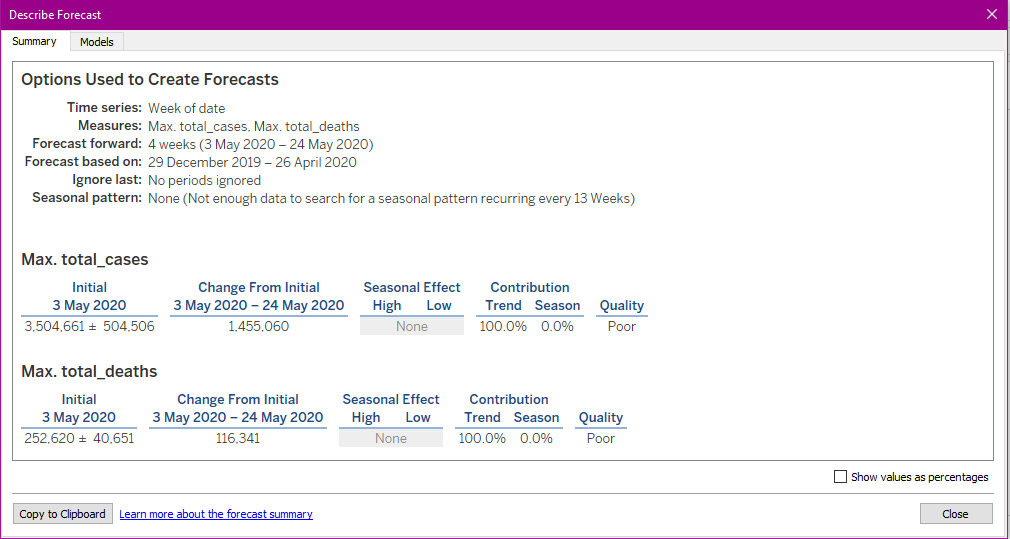
significance.

Tableau allows further analysis through forecasting the future values. By preparing a forecast based on model found suitable by Tableau for 4 weeks ahead we can observe the forecasted values for total deaths and total cases.

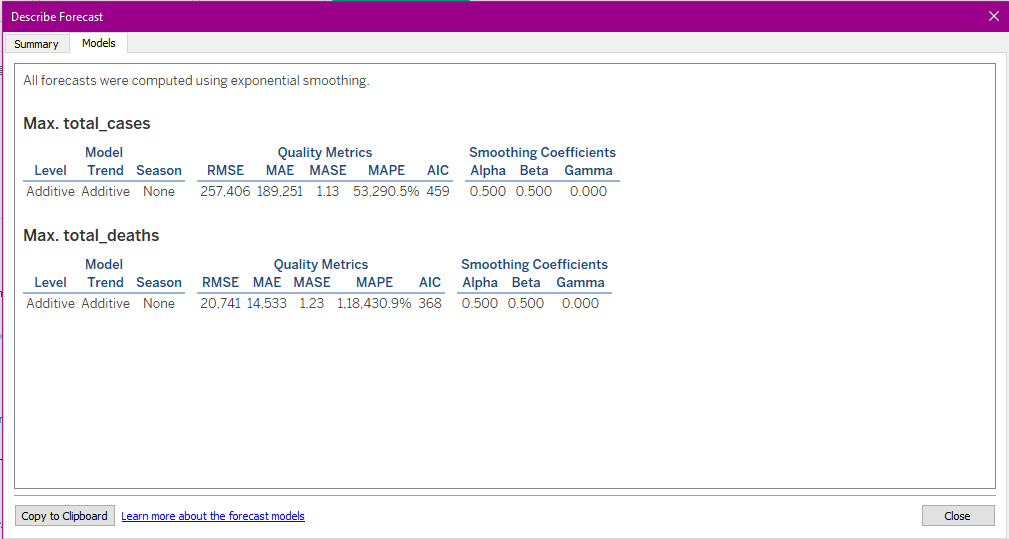


*Figure 19 Tableau Forecast options*

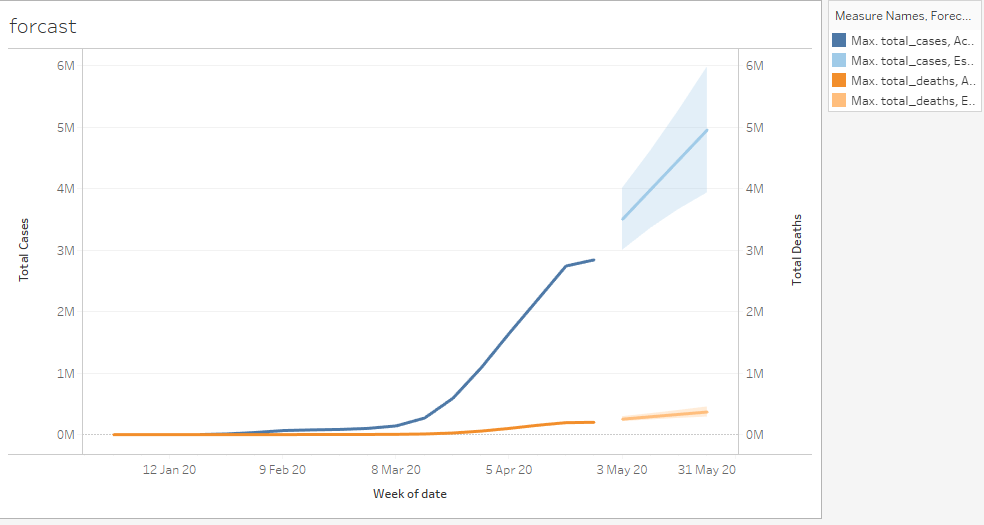
Description of forecast created for 3 May 2020 – 24 May 2020:



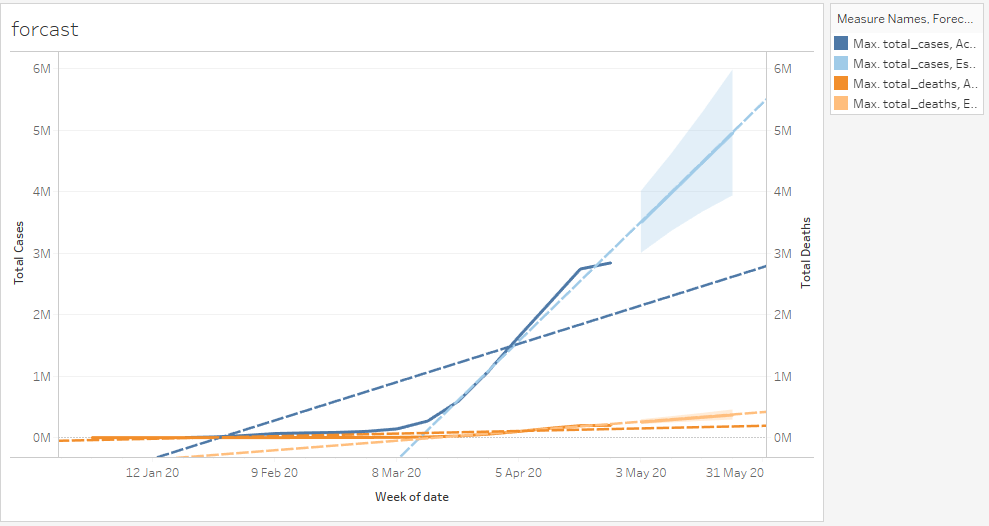
*Figure 20 description of forecast*



*Figure 21 Description of forecast model*



*Figure 22 Tableau forecast total deaths and total cases*



*Figure 23 Tableau forecast with trend lines*

Forecasting in Tableau uses a technique known as exponential smoothing. Forecast algorithms try to find a regular pattern in future-continuable steps. Tableau selects the best of up to eight models automatically, the best being the one which generates the highest quality forecast.

We can observe that both total deaths and total cases are increasing linearly in the forecasted values. It is further analysed if the model is reliable by comparison between forecasted values and real values.(Ourworldindata 2020) maintains data for coronavirus spread and is free to use for any purposes. Data for 3-May to 24 May was compared with forecasted values: -

|  |  |  |  |
| --- | --- | --- | --- |
| **Measures** | **10-May-2020** | **17-May-2020** | **24-May-2020** |
| **Total Deaths** | |  |  | | --- | --- | | Forecasted Value | Actual Value | | 3989691 | 3983189 | | |  |  | | --- | --- | | Forecasted Value | Actual Value | | 4474701 | 4586782 | | |  |  | | --- | --- | | Forecasted Value | Actual Value | | 4959721 | 5265120 | |
| **Total Cases** | |  |  | | --- | --- | | Forecasted Value | Actual Value | | 291401 | 278864 | | |  |  | | --- | --- | | Forecasted Value | Actual Value | | 330,181 | 311195 | | |  |  | | --- | --- | | Forecasted Value | Actual Value | | 368,961 | 341249 | |

*Figure 24 Table Actual vs Predicted*

Actual values for total deaths are very close to predicted values and are increasing at a linear rate. Therefore, the model can be useful to predict future values of reported deaths.

The total cases are predicted correctly for 10-May but as the end of month approaches, we can see the actual values are increasing rather rapidly than forecasted values which implies that the **total coronavirus cases are increasing exponentially rather than linearly**.

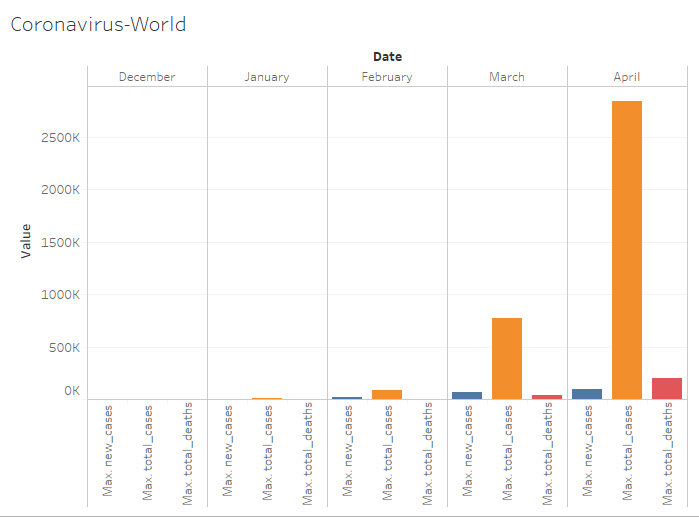
**Bar Plots**

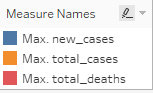
A bar plot shows the relationship between a numeric and a categoric variable. Each entity of the categoric variable is represented as a bar (Data-To-Viz 2020). The bar size stands for its numerical value. Bar Plots compares values between various classes and shifts in track over time. They also work well for categorical data.

**Advantages: -** They are easy to prepare, easy to update and are good for small projects.

**Disadvantages:** - They are limited to X-axis and Y-axis displays

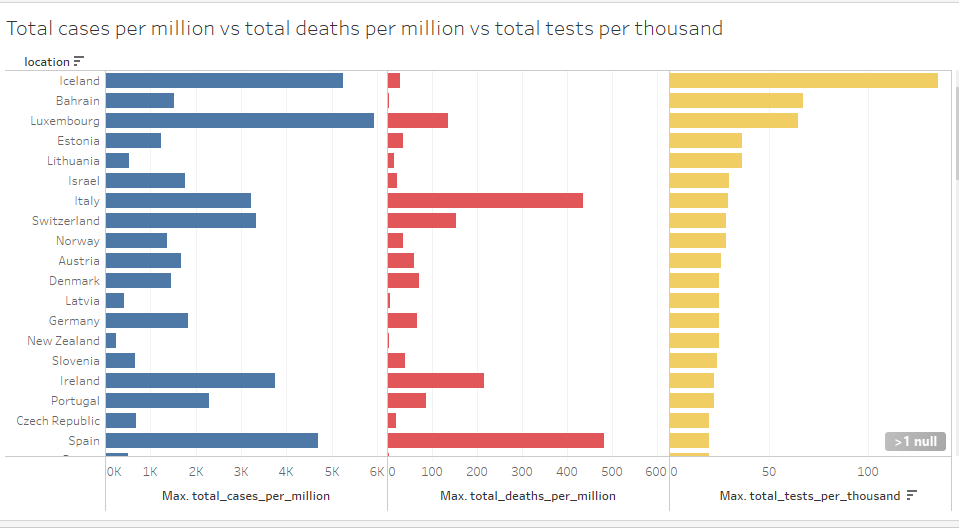
**Analysis:**





*Figure 25 World stats of new cases, total cases and total deaths*

This visualization shows very low total cases, new cases and total death cases in January and February, but a significant rise occurred in each type of cases in the month of March. However, **drastic exponential increase of 266% can be seen in confirmed cases while gradual increase in cured and death cases in the month of April**.



*Figure 26 total cases per million vs total deaths per million vs total tests per thousand*

To further analyse if reported coronavirus cases increased with tests performed a bar plot has been prepared with measures total cases per million vs total deaths per million vs total tests per thousand. Total cases per million correctly depicts the spread in the country as different countries have different size and tests per thousand also gives a proper measure to analyse the rate of tests performed in comparison to the population.

It is observed that though some countries which have high tests per thousand have high reported cases such as Iceland , there is **not any correlation of total tests with high cases** as there are countries were the numbers are opposite for example in Lithuania the test performed per thousand are high (36.6 ) with comparatively less cases per million (15.1).

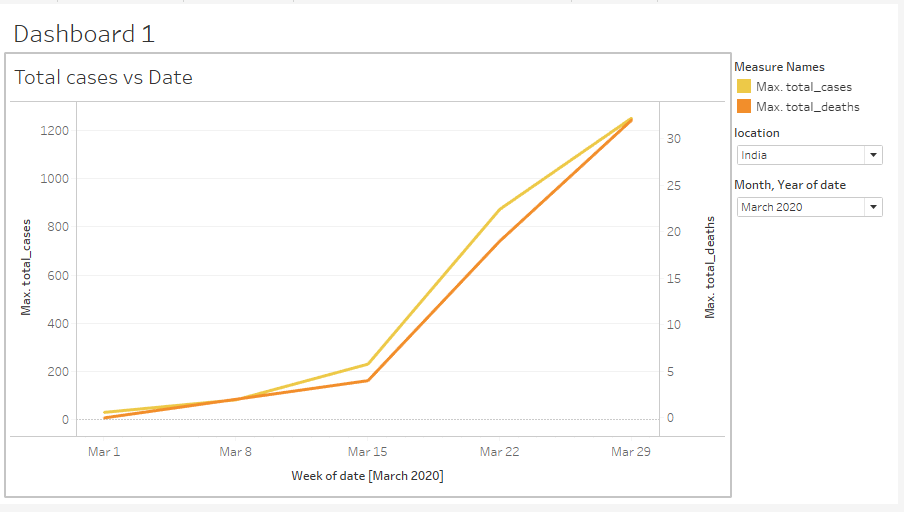
**Discussions**

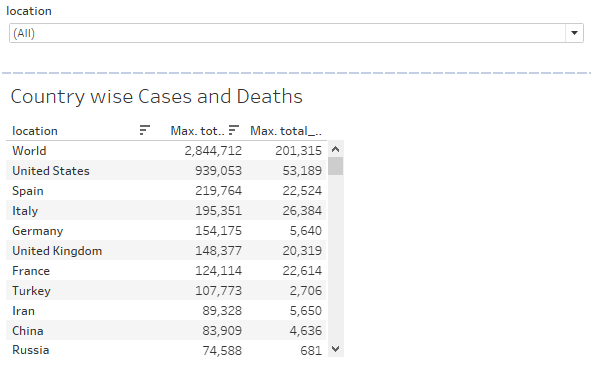
**Tableau Pros as observed: -**

There is a large community online which helps learning and implementing visualizations in Tableau very easy. The drag and drop functionality is easy to use and doesn’t require any programming knowledge.

**Tableau Cons as observed: -**

Proper planning is required to handle large number of dimensions and develop insights. It is not very strong in data preparation, modelling.

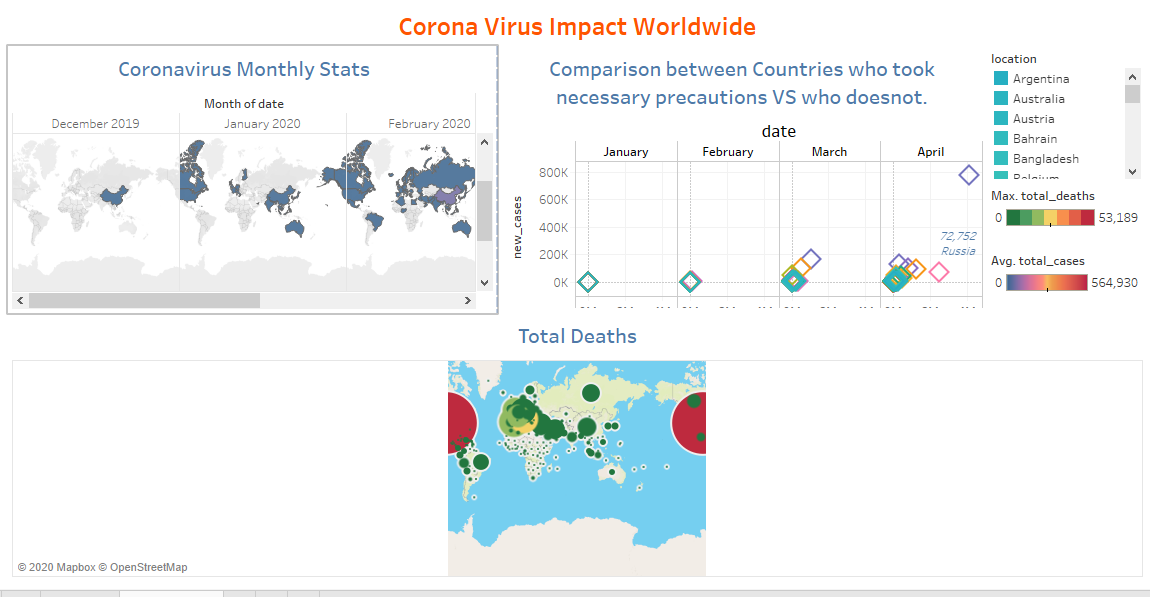




*Figure 29 Dashboard for total cases*

Dashboard is a user interface where information is structured and displayed in a way that is easy to read. The dashboard shown in Figure 29 is interactive in nature and makes tracking live updates and trends simpler. We can use drop down menu to select country and month to visualize total cases and deaths.

First sheet in dashboard shows line graph between death and cases. Country and Month can be selected from dropdown. Second sheet in dashboard shows coronavirus cases and deaths for each country, country can be filtered by selecting specific country from the dropdown.

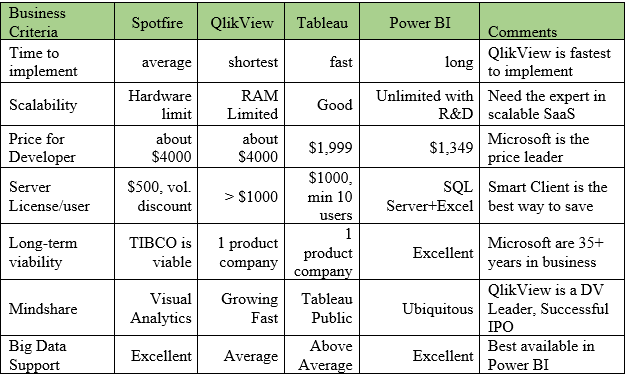


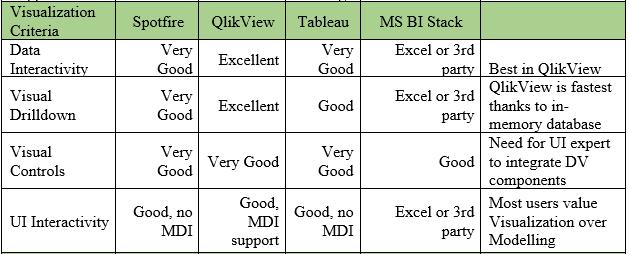
*Figure 30 Dashboard for corona impact worldwide*

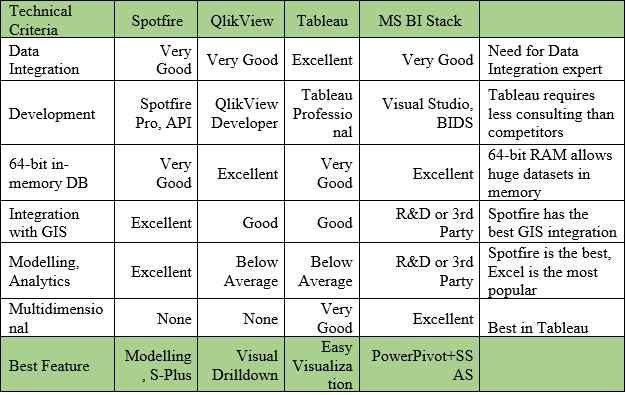
We can observe from the dashboard how this pandemic began from China and spread throughout the world within 3 months.

**Visualisation Tools for Multidimensional Data**

There are many tools available online for visualization of multidimensional datasets. Following table provides a companion between a few of them: -







**Conclusion:**

Multidimensional data visualisation involves the observation and analysis of more than one measure variable at a time. However, visualisations of this data can encounter problems related to the size and number of dimensions of the data. Careful planning and analysis should take place in order to generate useful graphs. This paper analysed multiple visualisation strategies to display the multidimensional data and showed that insightful research can be done using geospatial mapping, bar graphs, line charts, scatter graphs, and heat maps. In the present time, spread of covid-19 around the world has set off a spike of interest in data analytics and visualizations. This paper uses Tableau to generate high-impact visualisations of popular data analyses to help us see and interpret the data by using predictive analytics to enhance preferential covid-19 spread decision making.

**References:**

Akhtar, Nikhat & Tabassum, Nazia & Perwej, Dr.Asif & Perwej, Dr. Yusuf. (2020). Data analytics and visualization using Tableau utilitarian for COVID-19 (Coronavirus). Global Journal of Engineering and Technology Advances. Volume 3. Page 28-50.

Card, Stuart & Mackinlay, Jock & Shneiderman, Ben. (1999). Readings in Information Visualization: Using Vision To Think.

Data-To-Viz. (2020, 09 01). Bar Plot. Retrieved from Data-To-Viz:

https://www.data-to-viz.com/graph/barplot.html

Dzemyda, Gintautas & Kurasova, Olga. (2015). Multidimensional data visualization. Methods and applications.

Keim, Daniel & Hao, Ming & Dayal, Umeshwar & Janetzko, Halldor & Bak, Peter. (2010). Generalized Scatter Plots. Information Visualization. 9. 301-311.

Lexico. (2020, 09 01). Choropleth Map. Retrieved from Lexico:

https://www.lexico.com/definition/choropleth\_map

MacEachren, Alan & Kraak, J.A.. (2001). Research challenges in geovisualization. Cartography Geograph.. Inf. Sci. 28. 3-12.

Moon, Keon-Woong. (2016). Heatmap.

Ourworldindata. (2020, 09 01). Coronavirus Actual Values. Retrieved from Ourworldindata:

https://ourworldindata.org/coronavirus-source-data

Palmas, Gregorio. (2016). Visual Analysis of Multidimensional Data for Biomechanics and HCI.

Tableau. (2020, 09 01). What is Tableau. Retrieved from Tableau Software:

https://www.tableau.com/why-tableau/what-is-tableau

Wikipedia. (2020, 09 01). Line Chart. Retrieved from Wikipedia:

https://en.wikipedia.org/wiki/Line\_chart

Wikipedia. (2020, 09 01). Multidimensional Analysis. Retrieved from Wikipedia:

https://en.wikipedia.org/wiki/Multidimensional\_analysis

Worlddata. (2020, 09 01). Average Age of Countries. Retrieved from Worlddata: https://www.worlddata.info/average-age.php