# Analysis of Covid-19 data using visualisations

## Abstract

The main aim of this paper is to explore and gain a better understanding of visualisations at hand for multidimensional or hypervariate data. We will also conduct an analysis of covid-19 data, available from dec19 to Apr20, using some of the relevant visualisations and attempt to highlight any hidden trends and patterns within the data. The visualisation platform to be used for this analysis is Tableau.

## Introduction

Visualisation of multidimensional data is the presentation of 2D and/or 3D data using graphs, charts and pictures with a use of visualisation software or platform. By displaying multidimensional data in a graphical form, the analysis of data becomes easier as for human perception it is important that data is represented in low-dimensional space. The main objective of using visualisations is to develop some visual insights into the data set being analysed.

Thomas and Cook (2005) define visual analytics as the science of analytical reasoning facilitated by the visual interfaces of abstract data. Such techniques are quite helpful in identifying trends and patterns otherwise difficult to observe.

Ware (2012) gives a different view in his article by arguing that using visualisations not only allows the analysts to derive useful insights and patterns from the data effectively but also enables them to pick up the errors and data outliers easily.

This paper is divided into 3 parts: first we briefly introduce and describe the visualisation software used for visual analytics of given data, second we define the different visualisation tools and methods available and used for multidimensional data in this analysis, next we analyse the covid-19 data and discuss any hidden trends and patterns, and finally the results and conclusions derived from the analysis.

## Visualisation software

## **Tableau**

Tableau is an interactive data visualisation software introduced in 2016. It is considered a powerful and one of the fastest growing data visualisation tools in the business intelligence field. Its comfortable user interface and effective way of simplifying raw data into easily understandable format makes this software the number one choice for data analysts and experts. One of the major advantages of tableau is that it does not require any prior technical knowledge or programming skills to run and allows any non-technical user to easily make dashboard as per his need. The main features of this software are data blending, real time analysis, and collaboration of data.

## Visualisation methods

### **Bar charts**

A bar chart is a chart that presents categorical data in rectangular bars with length of each bar proportional to the value it represents in the data. The bars in a chart can be plotted vertically or horizontally. A vertical bar chart is commonly known as column chart. One of the main advantages of bar charts is its ability to represent data that shows changes over time, allowing users to discover trends easily.

### **Scatterplots**

A scatter plot is a data visualisation that displays the value of two different variables as points on a graph. The data for each point is represented by its horizontal (x-axis) and vertical (y-axis) position on the visualisation. It is mainly used to observe and show relationship between two numeric variables. Scatter plots allow the user to identify patterns and correlations easily. It also helps to visualise any clusters or groups and outliers in the given data. Usually the independent variable is plotted on the horizontal axes and the dependent variable on the y axis. the relationship between variables could be described as linear or non-linear, strong or weak, and positive or negative.

### **Geo visualisation**

Geo visualisation, short for geographic visualisation, deals specifically with displaying information that has a geospatial component to it. A geospatial component is geographic or positioning information. It is a process that alters geographic information to enable us to absorb it with our eyes. The purpose of this method of data representation is to capitalise on our affinity for visual things and convert the seemingly random collection of data available to us into something that is easily understood. According to statistics, approximately 85% of the information available today has some sort of a geo spatial component to it, which makes geo visualisation an important tool to represent the data in a visual form by using the geo spatial component of the data.

## Data

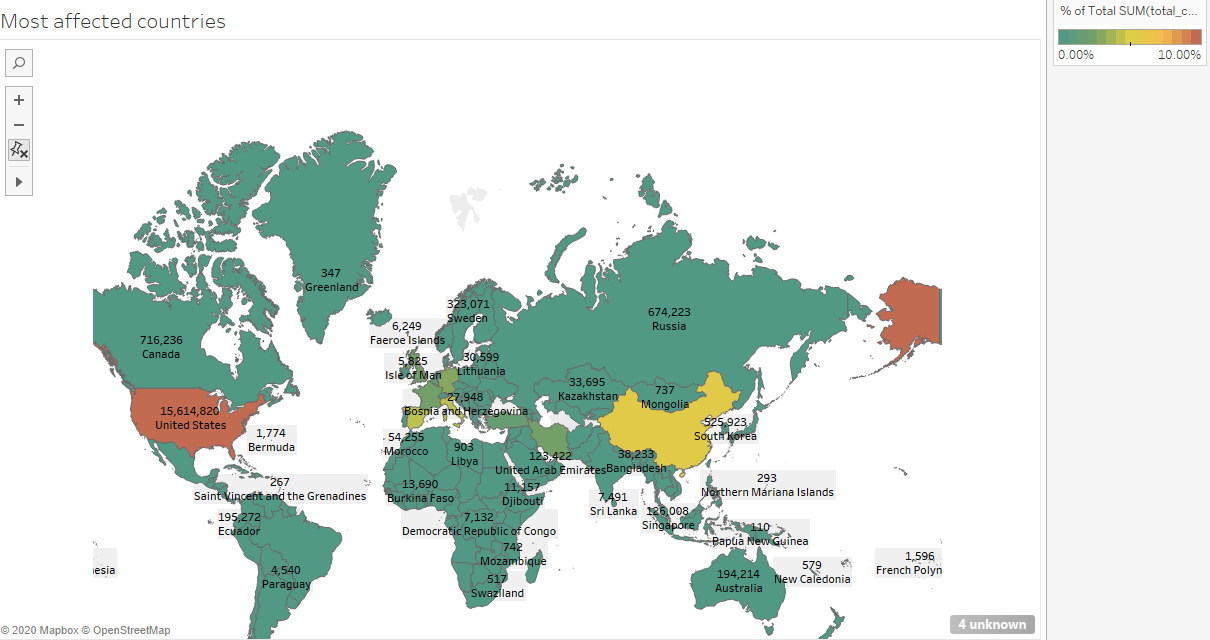
### Overview

The data set used for visual analysis in this paper is the Covid-19 pandemic data for the world. The data was retrieved from *Our World In Data* website and can be found at: <https://ourworldindata.org/coronavirus-source-data>. The covid data used for our visualisations is from January 2020 to end of April 2020. It is sorted by geographical location and has multiple measures including total cases, total tests performed, total deaths, new cases, new deaths, cases per million, and so on. Since we have a limitation in our data that it only reflects the pandemic situation around the globe for the first four months of the outspread, we will only be analysing and deducing the results from first four months of pandemic.

## Visualisations & Analysis

### Graphs

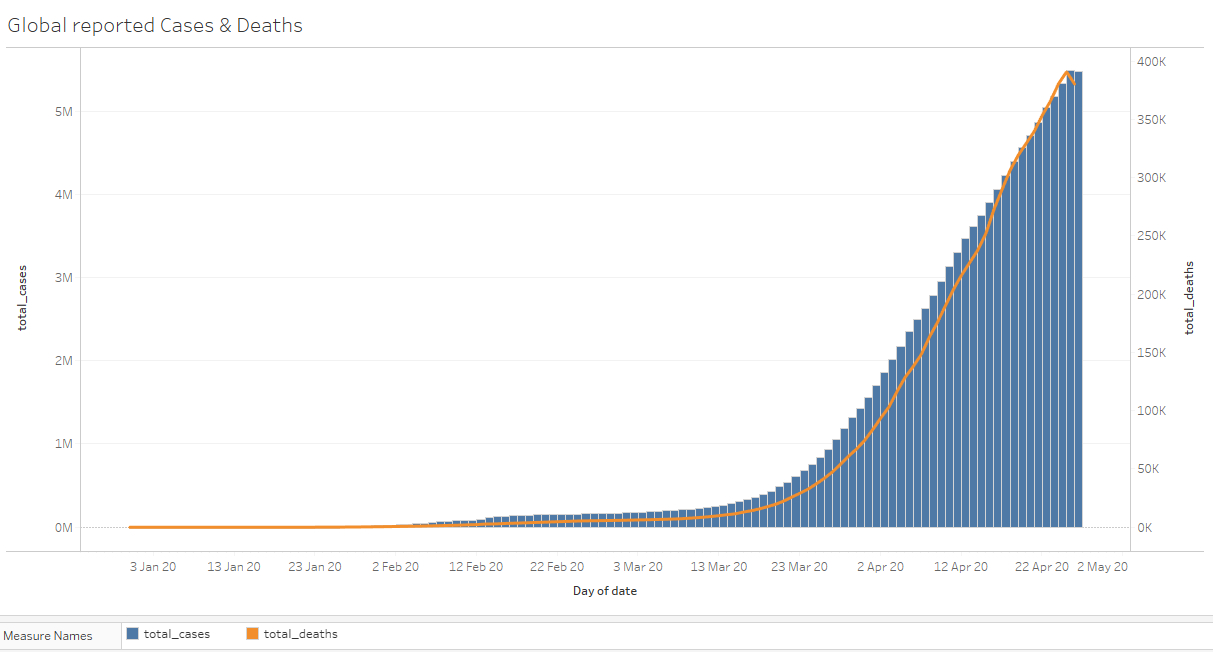
All the visualisations were generated using Tableau software. We first show the geo visualisation of the covid-19 data, highlighting most affected countries using colours. The diverging temperature palette was used, green colour showing least affected countries while red colour reflecting most affected countries.



From the map above, United States is the country most affected by Covid-19, followed by China, United Kingdom, Germany, France, Italy, Spain, Iran, Turkey, and Belgium. As a fact, we know that China was the first country to be affected from corona virus. In the first four months, the countries most affected by the virus are mostly from Europe and United States. This could mean that there were high international travels between US and China, and between European countries and China, due to which the virus spread to these countries quickly. The below map represents data according to size of cases in each country. The most visible countries are US, China, UK, Spain, Italy, Germany and France.

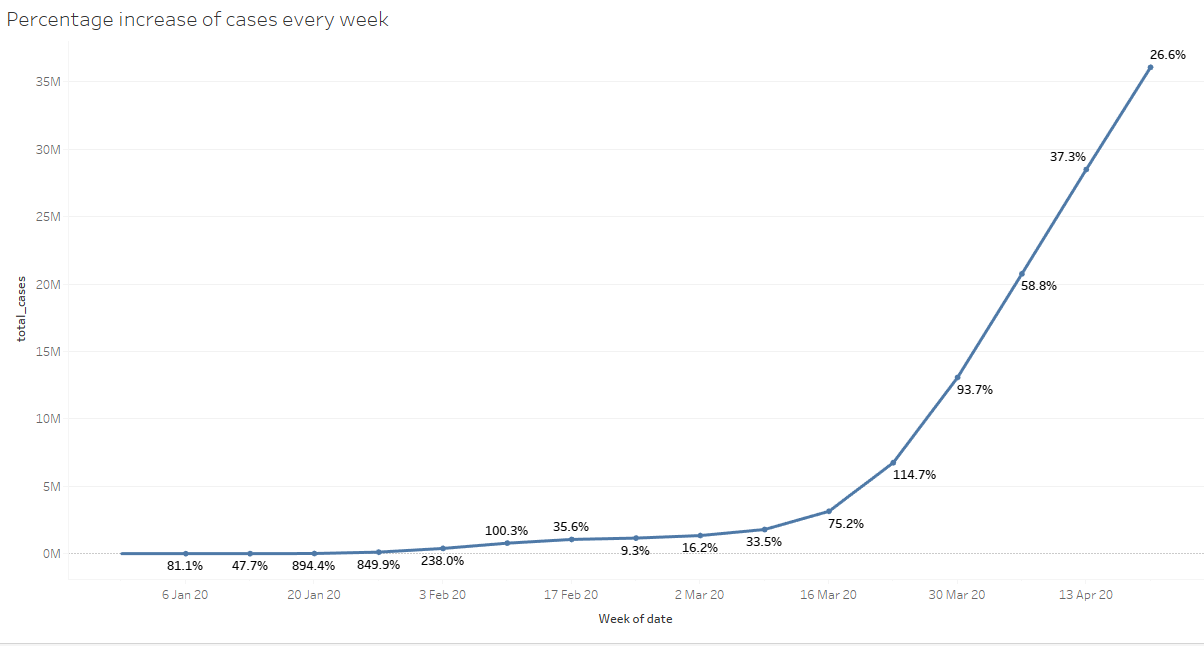


For the next visualisation, we used a combination of bar chart and line graph to show the total global cases and deaths reported from January to April 2020. We used dual axis for this visualisation to enable easy interpretation of data and to allow us to observe any trends.



from the above bar chart, it is observed that the total reported cases for the first 80 days were only less than half a million which rapidly increased in the next one week to above one million cases. The graph takes a dramatic rise after 23 march and becomes steep for the next 30 days. If you look at the graph closely, you will notice that the total number of deaths represented by a line graph follow the same pattern. The deaths reported grow sharply from start of April. At the end of April there were more than 5.5 million total corona cases reported all over the world with around 0.4 million deaths from the pandemic. Therefore, it will be interesting to see the actual weekly percentage increase in total cases over the span of first four years and how has the testing affected the number of total cases being reported.

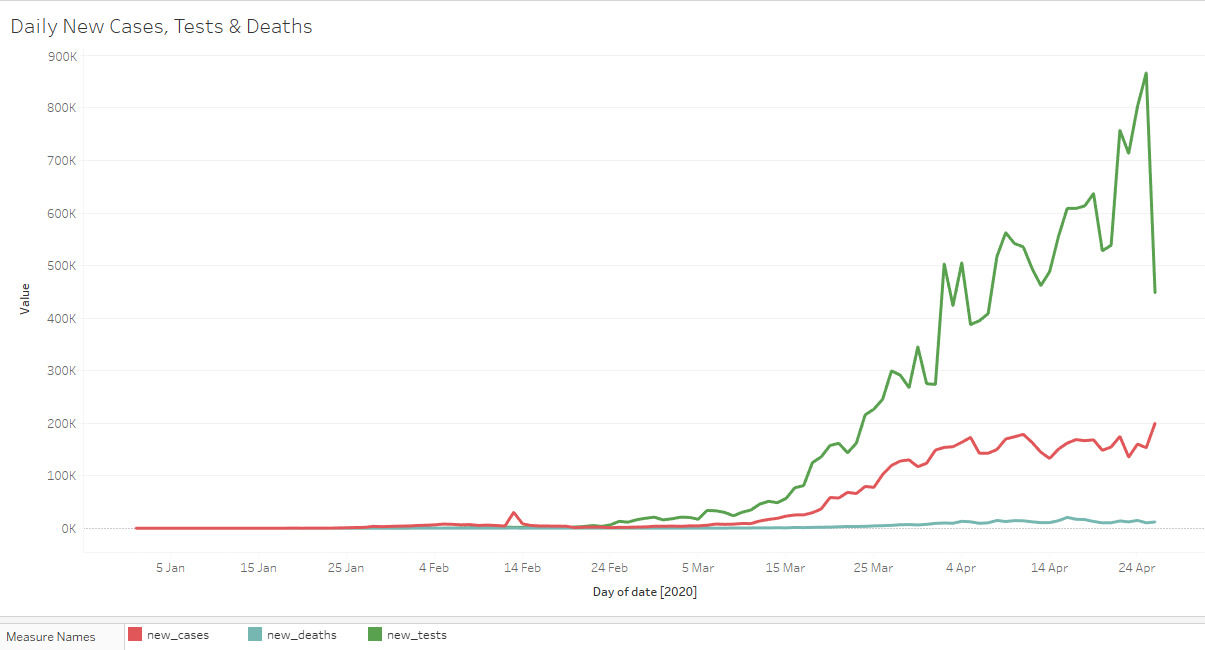
To visually analyse this, we take help from the line graph and plot total cases against date as the dimension. We use quick table calculation option and use percentage difference which calculates each current value as a percentage difference from the previous value. Moreover, we used label text from the marks card to show the weekly percentage increase on the line graph.



If we look at the line chart above, we can see that since that there’s a massive increase in percentage of total cases accounting to 894% from the previous week. This is where the cases reported jumped from 1220 to 12,132 in one week and then from 12,312 to 115,246 in the following week. The percentage increase in cases slowed down for the next few weeks before the cases boom rapidly again in the weeks 11, 12, 13, and 14 of the data. The percentage increases calculated were 75.2%, 114.7%, 93.7%, and 58.8% respectively.

Now it would be worth observing how testing has affected the increase in total cases reported over the span of four months. For this, we use line charts and symbol maps to represent the data.

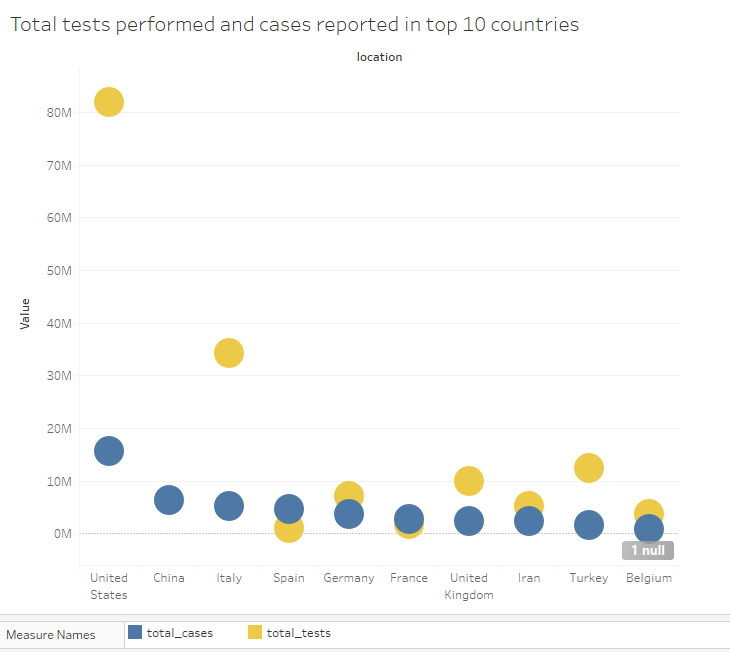
First, we use line charts to show display the daily testing performed globally and compare it in the same chart with new cases and new deaths reported each day.



From the above, we can see that the daily tests performed, and the new cases reported arguably remained low from January to the end of Feb, after which the daily testings were increased, and the new cases being reported also started to rise. There is no steady increase in the number of cases reported or the daily tests performed as we can see that there no straight lines in the above graph. We can still say that the number of new cases reported each day increased as the number of daily tests performed increased. In the last part of the graph, the difference between the cases and tests performed grows larger. The biggest difference observed was on 25th of April where the tests performed were 866,947 while the new cases reported the same day were only 153,214, a difference of more than 700,000.

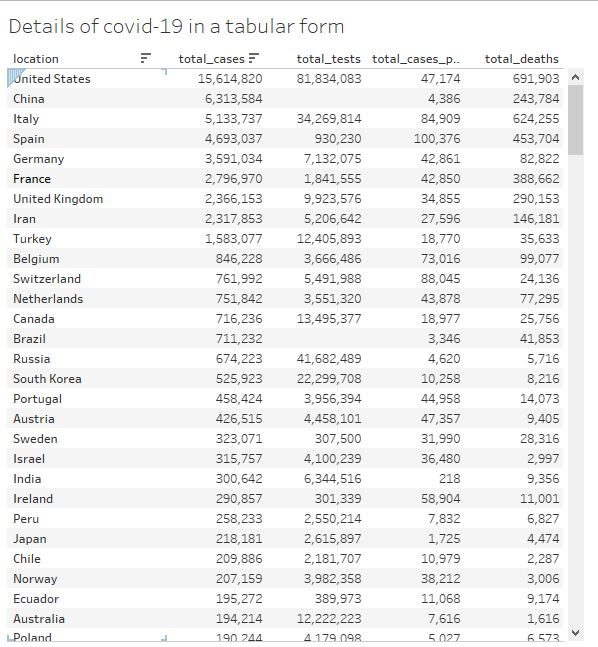
Now we will visualise the data of top 10 countries according to total cases reported and tests performed during January and April.

For the below visualisation, we used circles on a chart to represent data and colours to differentiate the cases reported against the tests performed.



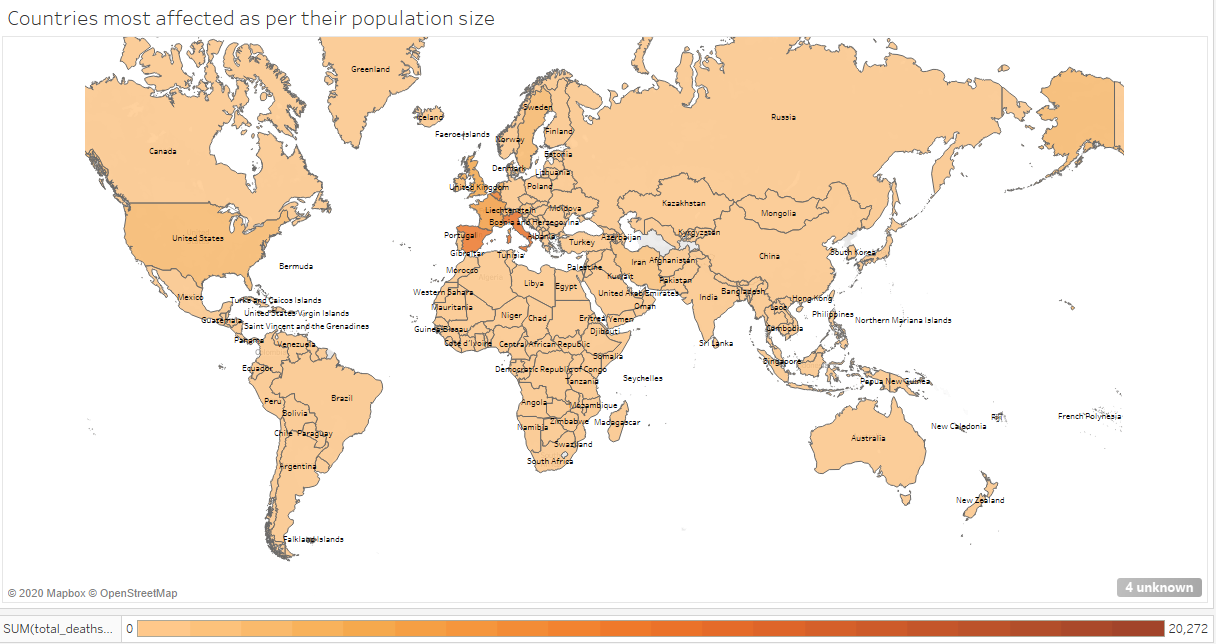
The graph shows that US performed most tests in the first four months of the pandemic (81,834,083). Other countries to perform aggressive testing were Italy, Turkey, and United Kingdom. There was no testing data available for China. It is also noticeable that Spain and France reported more cases than the tests they performed. There was small difference for Belgium and Iran between the two variables.

For easier comparison of cases and tests between countries, we also represented the data in a tabular form with additional columns showing cases per million population in each country, deaths per million, and tests performed per thousand.



We sorted the table in a descending order with countries with most cases on top. From the above table, we can deduce that overall US is the most affected country with highest covid cases of more than 15 million and almost 0.7 million people losing their lives in the first four months of the pandemic. China and Italy follow the list with more than 5 million cases each and high number of deaths reported as well. Although, tables provide easy comparison across rows and columns, its difficult to observe which countries more reported most cases and deaths per million population.

To represent this data, we will make use of geo visualisations again to show on a map which countries were most affected as per their population.



Using this method makes it easier for users to pick up the areas on the map most affected as per their population size. From the map we can deduce that Spain seems to be affected most by the corona virus with 100,376 cases reported and 9,700 deaths. Italy and United Kingdom are other major countries affected followed by USA. Surprisingly Sweden also shows a high death to cases per million ratio, a hidden observation which would have been ignored in other visuals. Geo visualisation method proves its worth in visual analytics through this hidden observation.

## Conclusion

The visual analysis of the covid-19 data shows that USA, Italy and Spain have been most affected by the covid-19 pandemic. US and Italy have taken serious steps towards testing of the virus by aggressively testing for the disease and ensuring that most cases are reported officially. Spain, Italy and Sweden have most cases per their population size from which can deduce that the virus spread was fast in these countries and their governments were slow or ineffective in taking initial steps towards containing the spread of virus.

To conclude, we can enhance analysis of multidimensional data by using visual analytics and choosing the right visualisation techniques and tools. Visual analytics allows experts to do a detailed analysis of complex data by representing data in a lower dimension. This approach also helps to spot trends and patterns through human eye, which would otherwise be difficult to observe with traditional techniques using algorithms and machine learning.

## References

Ware, C 2012, Interactive Technologies : Information Visualisation : Perception for Design (3rd Edition), Morgan Kaufmann Burlington, MA, USA.

Thomas, JJ, & Cook, KA 2005, Illuminating The Path : The Research and Development agenda for visual analytics, IEEE Computer Society Press, Los Alamitos.

Tableau, <https://www.tableau.com/why-tableau/what-is-tableau>

Our World In Data, https://ourworldindata.org/coronavirus-source-data