|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Data Visualization Project  Tableau Dashboard Part - 2 |
| |  |  |  | | --- | --- | --- | | Reonne Benoy – A00268832 | 4/27/20 | Data Visualization | |

Contents

[Dataset information 2](#_Toc38885512)

[Analysis of Visualization 2](#_Toc38885513)

[Final Dashboard 2](#_Toc38885514)

[Questions and Solutions Visualization 3](#_Toc38885515)

[1. Represent the Covid-19 pandemic effect and display the progression. What are the conclusions that could be taken from the time series progression and show how the effect projects in the near future? Give a detailed explanation. 3](#_Toc38885516)

[2. Display the top cases and the countries which are affected by the pandemic using an appropriate graph. Use statistical measures to find the confidence interval and the average of the confirmed cases. Use a condition to filter the number of countries where the cases are greater than 10000. Would using a cluster classification be beneficial? 6](#_Toc38885517)

[3. Display the variation in the number of cases in the top countries that are affected by the Covid-19. 8](#_Toc38885518)

[4. Use one of the data visualizations features to represent the global pandemic with perceptual accuracy. State why the feature is relatively better. 11](#_Toc38885519)

[5. Is the data influenced by the presence of outliers and if so, how far is it influenced? Take an example of top 3 countries with highest cases and prove the same. 13](#_Toc38885520)

[References 15](#_Toc38885521)

Tableau Dashboard

Data Visualization

Dataset- Covid-19.csv

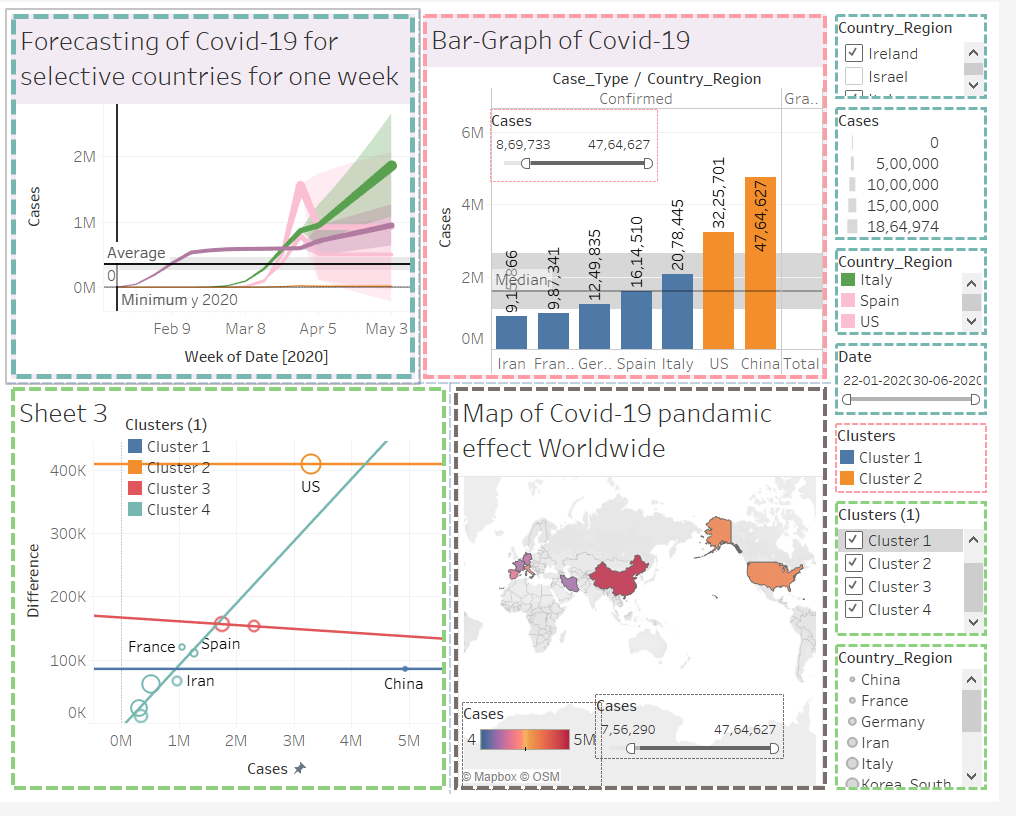
# Dataset information

COVID-19 case data can be directly downloaded or accessed through a Web Data Connector from data.world. Refreshed daily by 10 p.m. PT

|  |  |
| --- | --- |
| Case\_Type | Confirmed or Death |
| Cases | Number of Cases |
| Diference | Diference between confirmed or death cases from the previous update. |
| Date | Date of recording the data |
| Country\_Region | Country and region names |
| Province\_State | Province and state names |
| Latitudes | Latitude of the location |
| Longitude | Longitude of the location |
| Table\_Name | Time series table name |
| Prep\_Flow\_Runtime | Time and date of recording of data in the database |

# Analysis of Visualization

## Final Dashboard



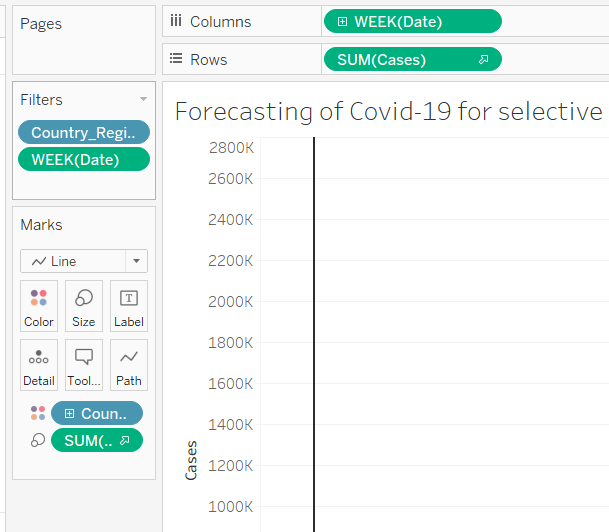
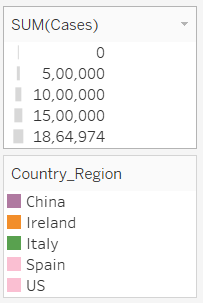
The above figure is the final dashboard with four different plots and information. Each plot is highlighted with a different layout and its corresponding filter is highlighted with the same.

# Questions and Solutions Visualization

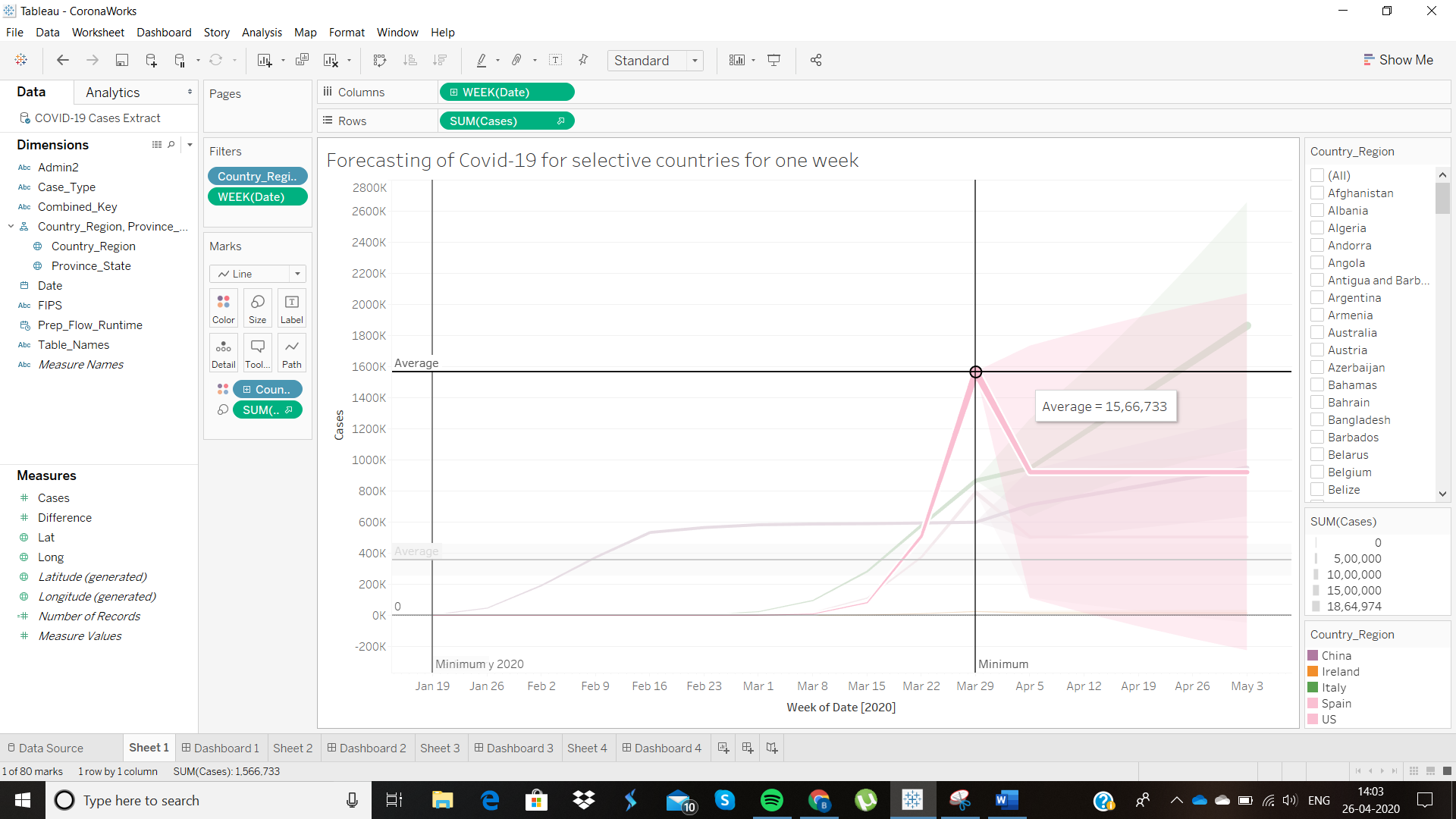
### Represent the Covid-19 pandemic effect and display the progression. What are the conclusions that could be taken from the time series progression and show how the effect projects in the near future? Give a detailed explanation.

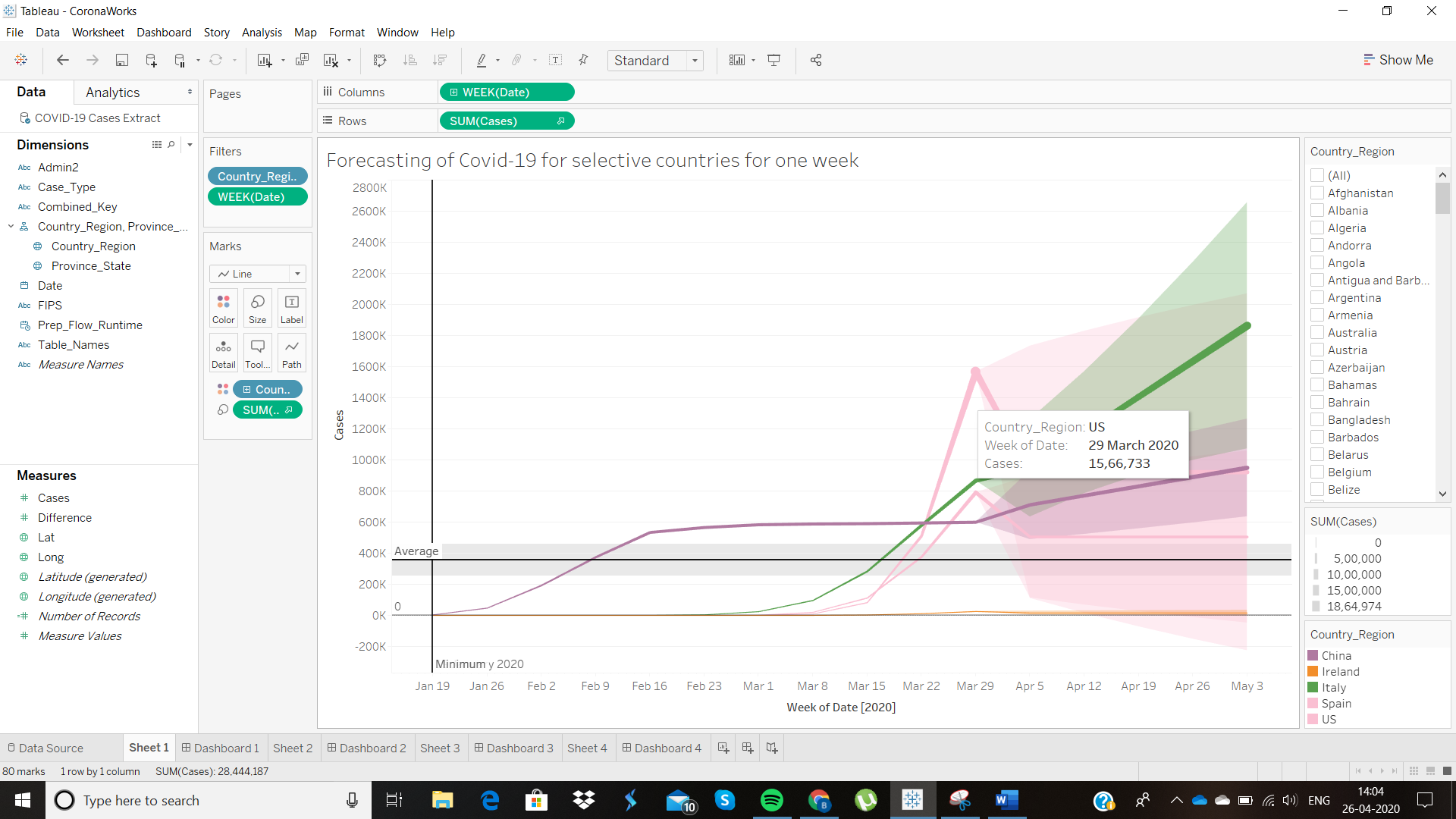
**Solution:**

Time series and progression can be visualized through a line graph. For a line graph, we need date and numerical data for visualization. We use the Date data represented through Weeks and the number of cases. We also use Country\_Region as a filter along with dates for data navigation.

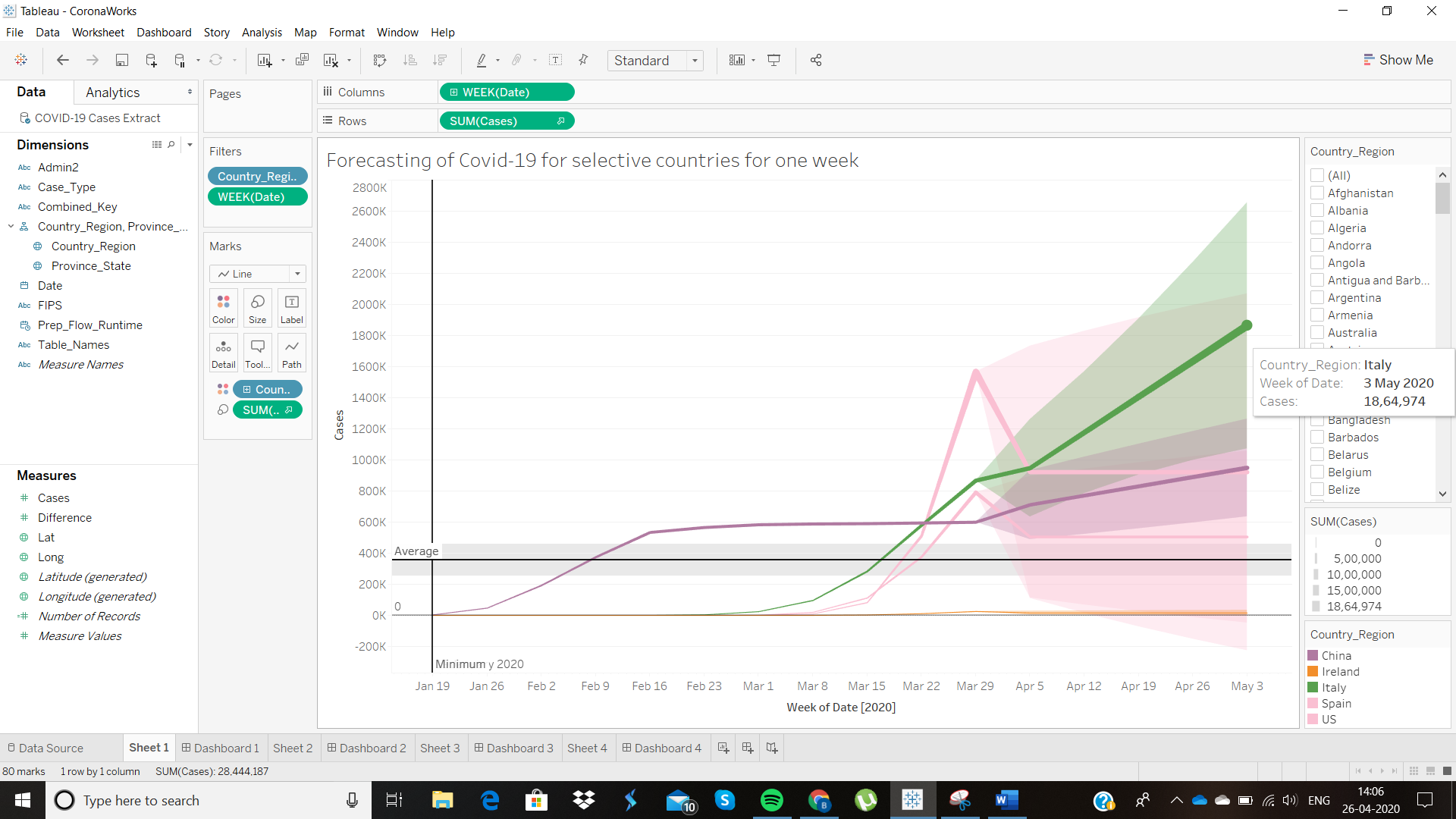
 

Creation of the line plot shows how the virus has spread, how the countries are affected and how it could possibly be affected in the future (forecasting). Below figure shows the line graph for US and the average population of confirmed cases.





Forecasting is used to project how the values will change in the near future. It depends on the data and the accuracy for it to be effective. Since this data begins from January and ends around April week 1, there isn’t many weeks of data for the forecasting to be completely accurate. Below we see that the projection of confirmed cases for Italy as of May 3rd will be close to 2 million.



|  |
| --- |
|  |

The following are the other statistical deductions which could be made from the data. We get the Confidence intervals for 95% confidence level and the average values.

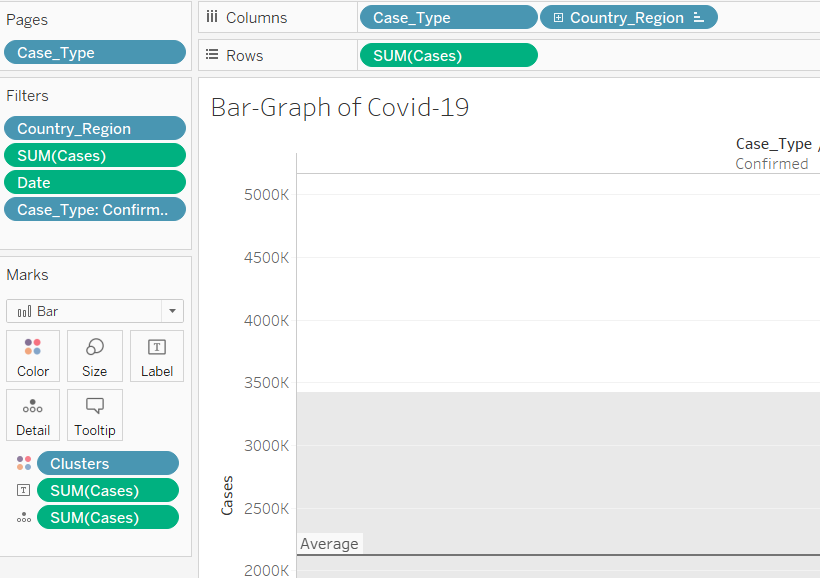


The forecasting of the effects of Covid-19 for 5 different countries can be clearly seen on the dashboard. Hovering over the line graph, we see the details of the progression projected over weeks for months, starting from January to April. We add projection for a week from April 26th to May 3rd displaying the forecast. From the forecast, we see that the number of cases in Italy projects to be highest by May 3rd. There are high chances where the forecast might be incorrect as the case data are of only four months. Hence, we forecast the readings for a projection on a linegraph.

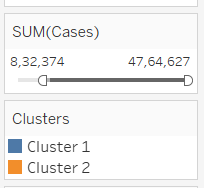
### Display the top cases and the countries which are affected by the pandemic using an appropriate graph. Use statistical measures to find the confidence interval and the average of the confirmed cases. Use a condition to filter the number of countries where the cases are greater than 10000. Would using a cluster classification be beneficial?

**Solution:**

For a bar-graph, we need to use one categorical variable and one numerical variable at the very least. For our example, we use two categorical variables, Case\_Type and Country\_Region. The Case-Type attribute denotes the confirmed cases and death whereas Country\_Region attribute denotes the country names, regions and provinces.



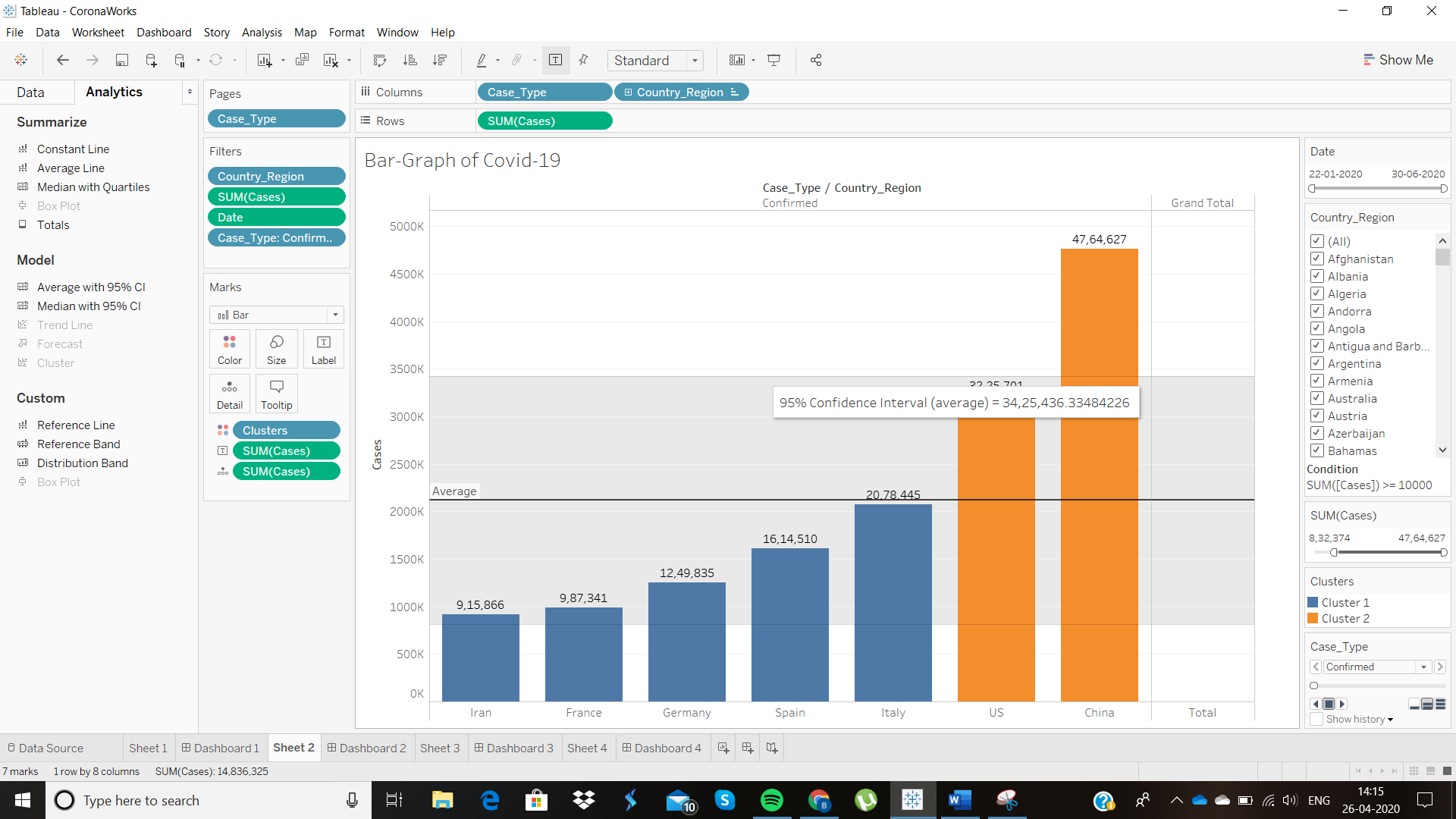
We use several filters such as Date, cases, Case\_Type and Country\_Region as shown above. We also try to implement a cluster classification to show how the effect of the Covid-19 are similar in countries at present.



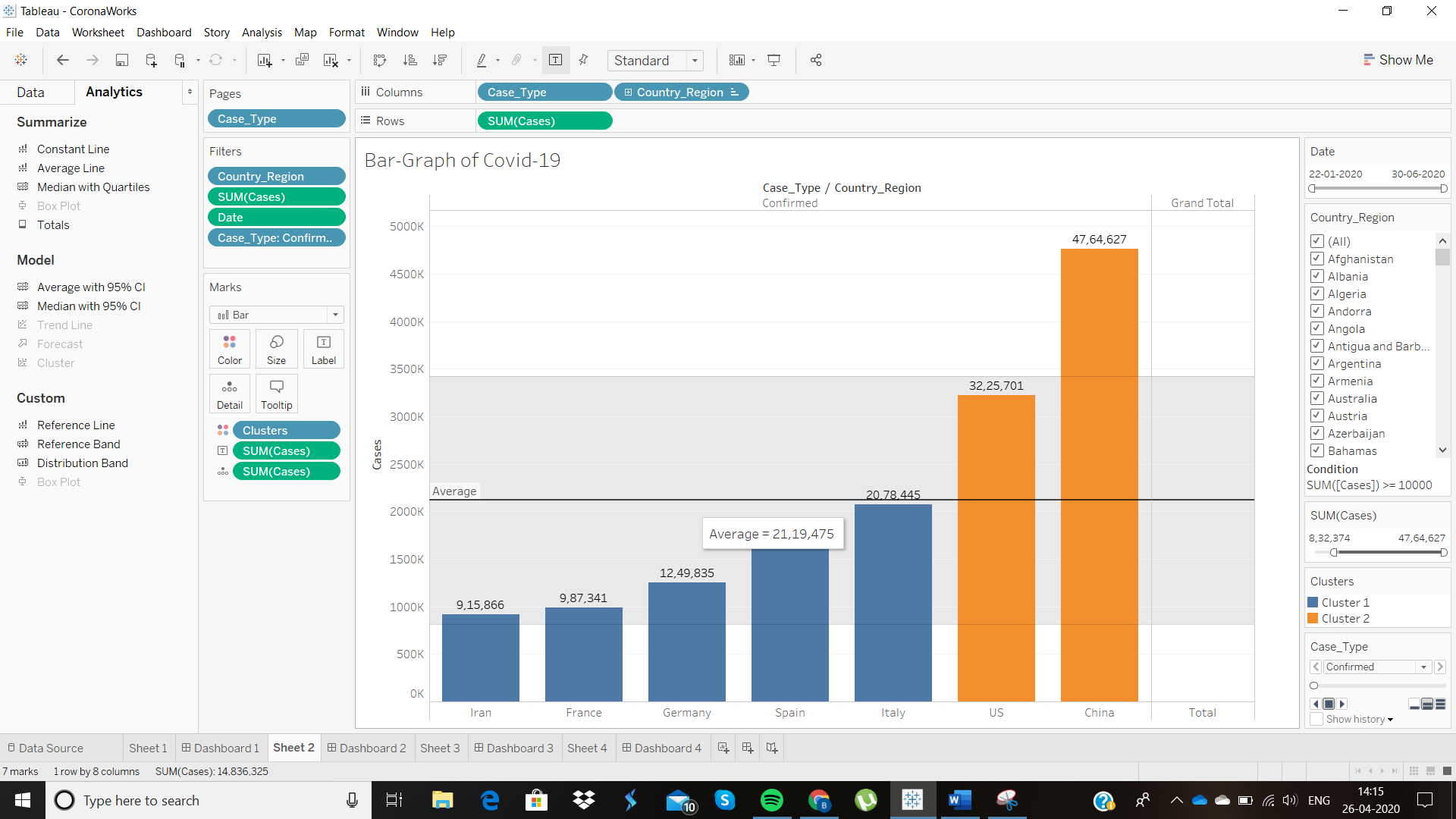
In the above figure, we use a filter where the sum of cases is greater than 800,000 at least. Also, prior to that, we use a condition where the sum of cases is greater than or equal to 10000.

We use the bar-graph to find the confidence interval for 95% confidence level based on the mean. However, we could also find the confidence interval based on the median as well.

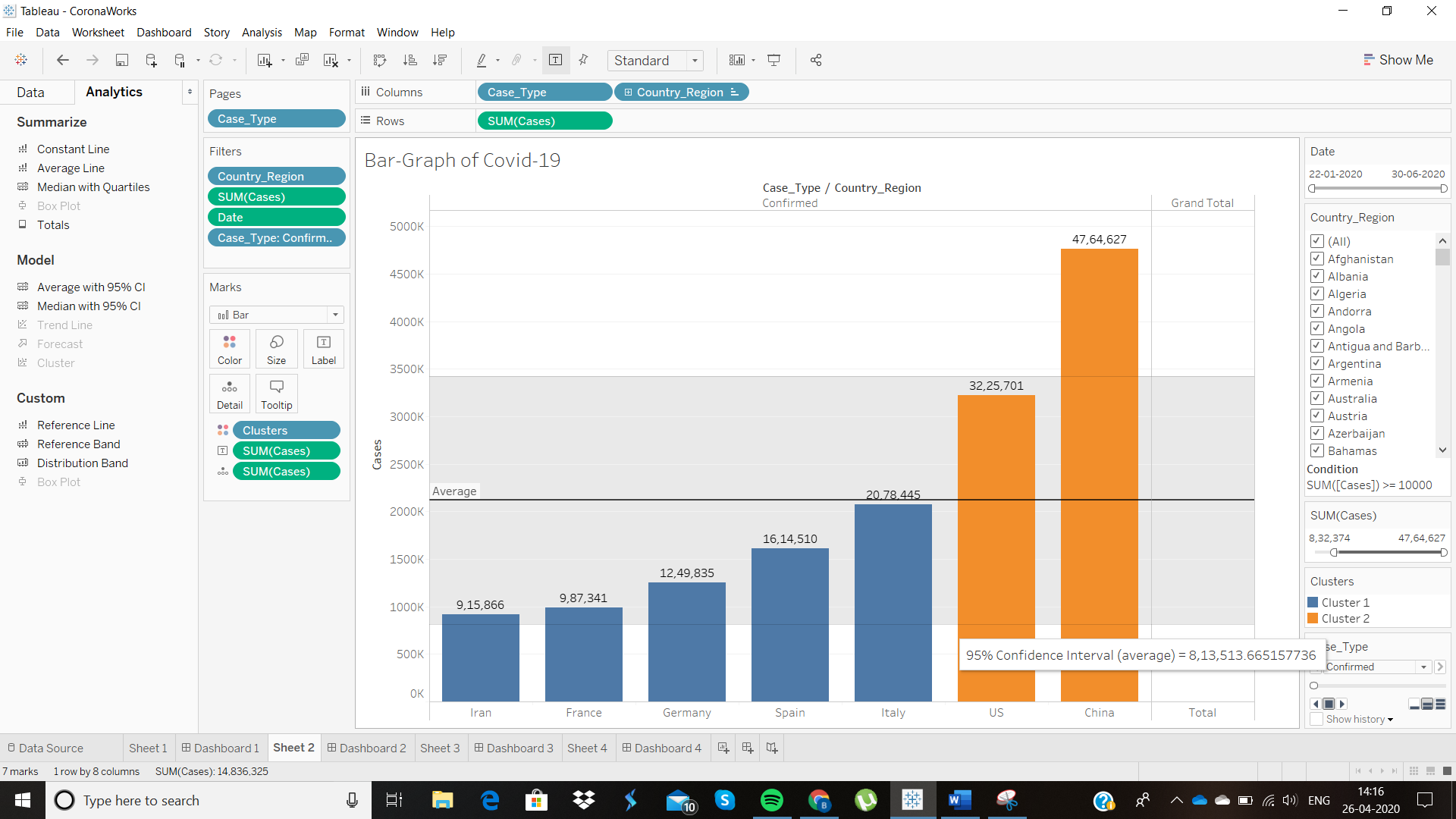
The following figures show the confidence interval for 95% confidence level based on the average value.



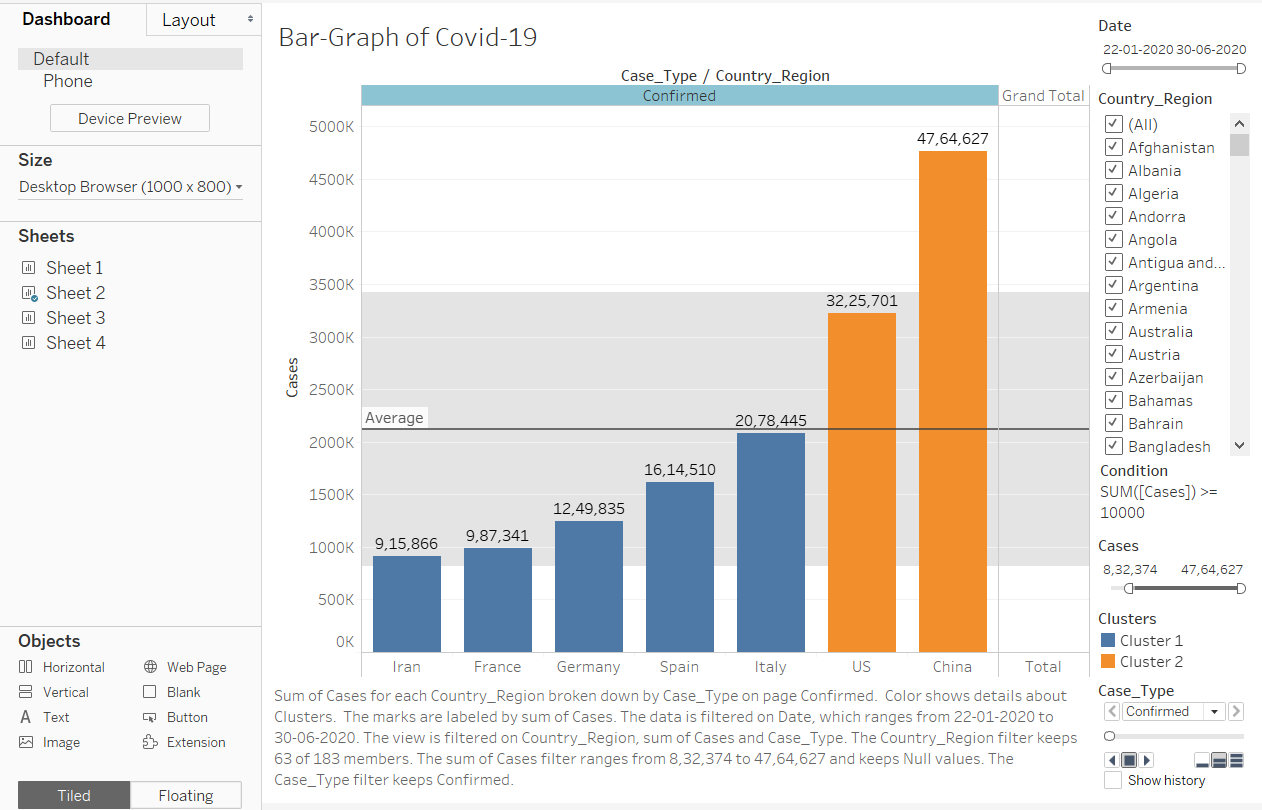
We get the average of the cases for 7 countries as shown below. The average is shown to be around 2 million.



Below is the lower band confidence interval for 95% confidence level.



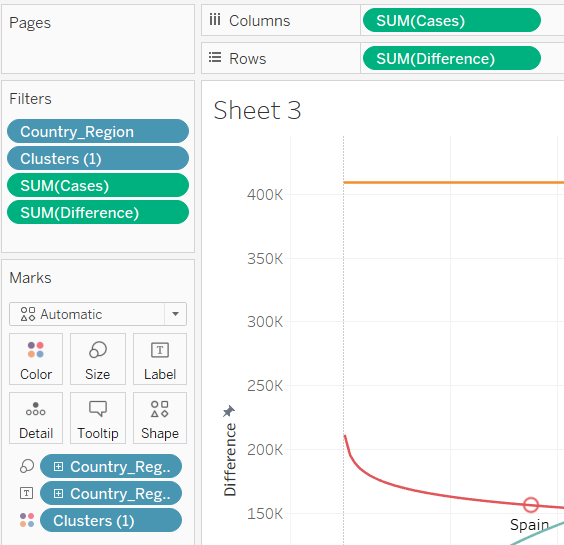
Creating a dashboard makes is easier for display and is better for visualisation of the final results. We also see that we use cluster classification of the data attributes based on the cases. We use two clusters and there is an option to use a greater number of clusters. We can deduce, from this information that, the population affected by the global pandemic and where the confirmed cases stand globally. From the dashboard, we see that Cluster 1 has Iran, France, Germany, Spain, and Italy whereas Cluster 2 has China and US.



### Display the variation in the number of cases in the top countries that are affected by the Covid-19.

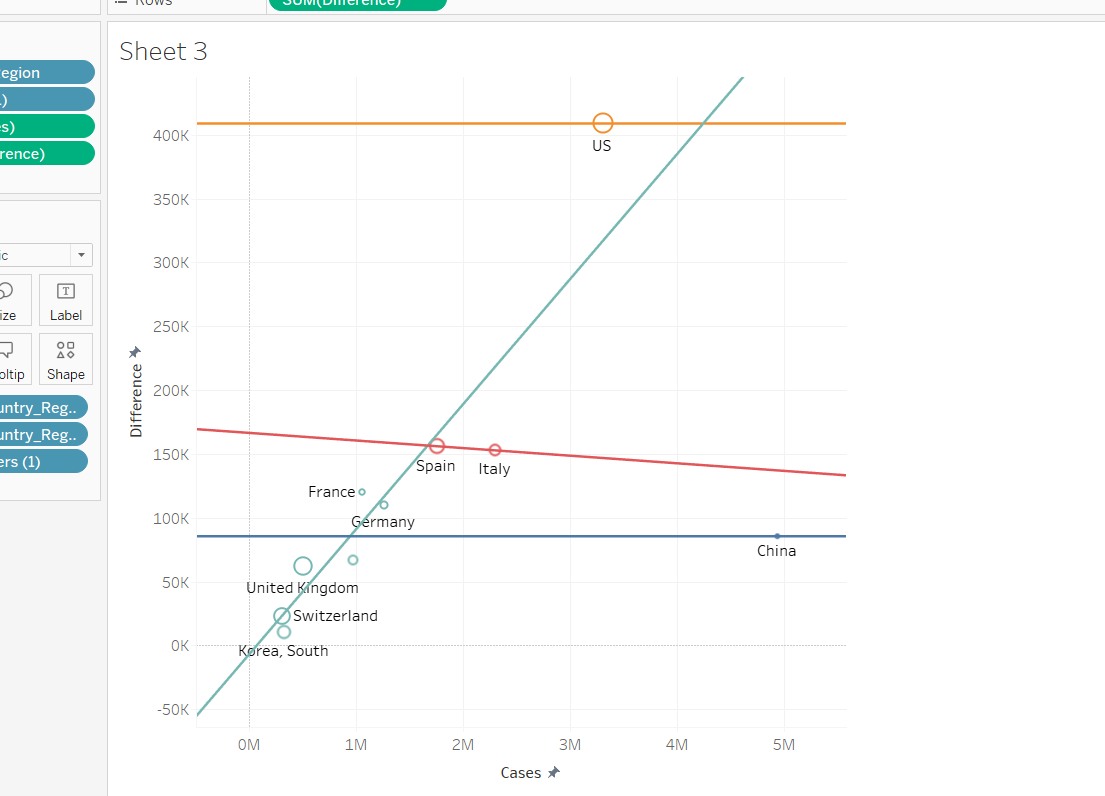
**Solution:**

Scatter plots need two or more of the continuous variables, therefore we use the number of cases and the difference attributes. We add filters with the number of cases, the differences and Country\_Region as shown below.

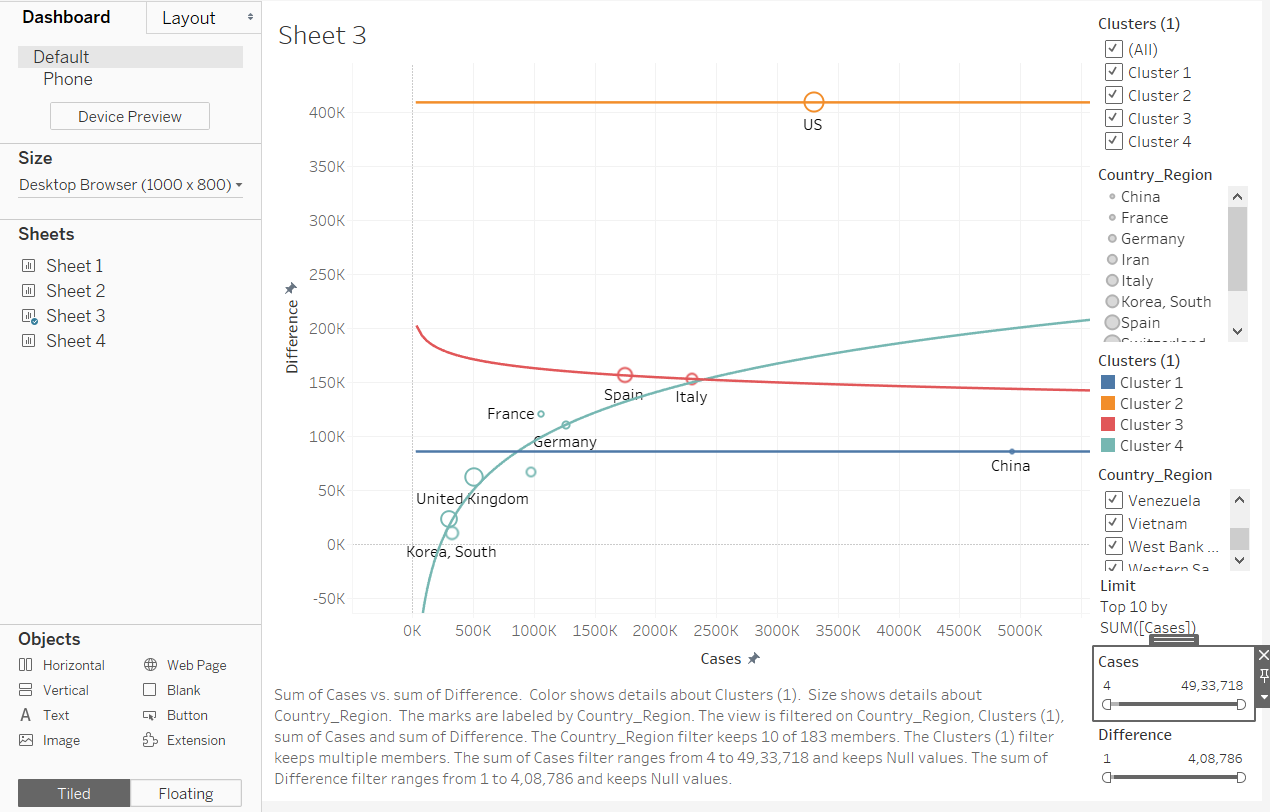
 

|  |
| --- |
|  |

From the above graphs, we see the variation in the number of cases. We use a logarithmic graph to see the rate of change of cases. From the graph, we see that one curve has a p-value which falls below the 0.05 threshold.



The above scatter plot shows a linear trend line on the graph. The linear trend line is used to show the change in the cases and how afar it is from the difference in the cases previously recorded.

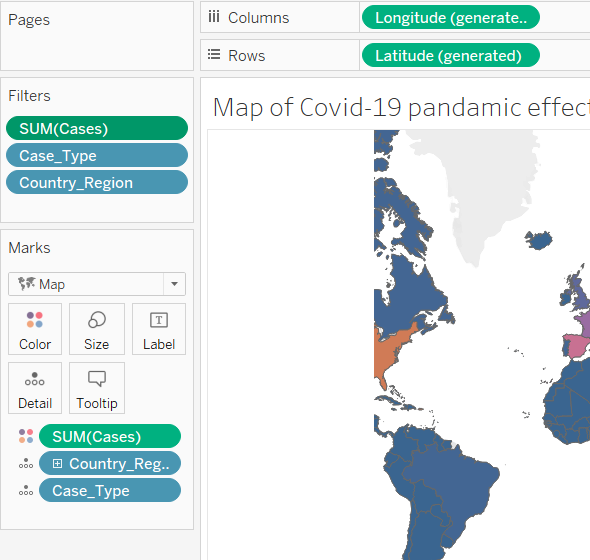
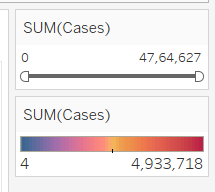


For our question, we deal with the logarithmic graph because only with the logarithmic graph, are we able to find the variation in the number of cases globally. Logarithmic graph deals with the rate of change of a value of the variable and hence, the variation in the number of cases can be visually represented.

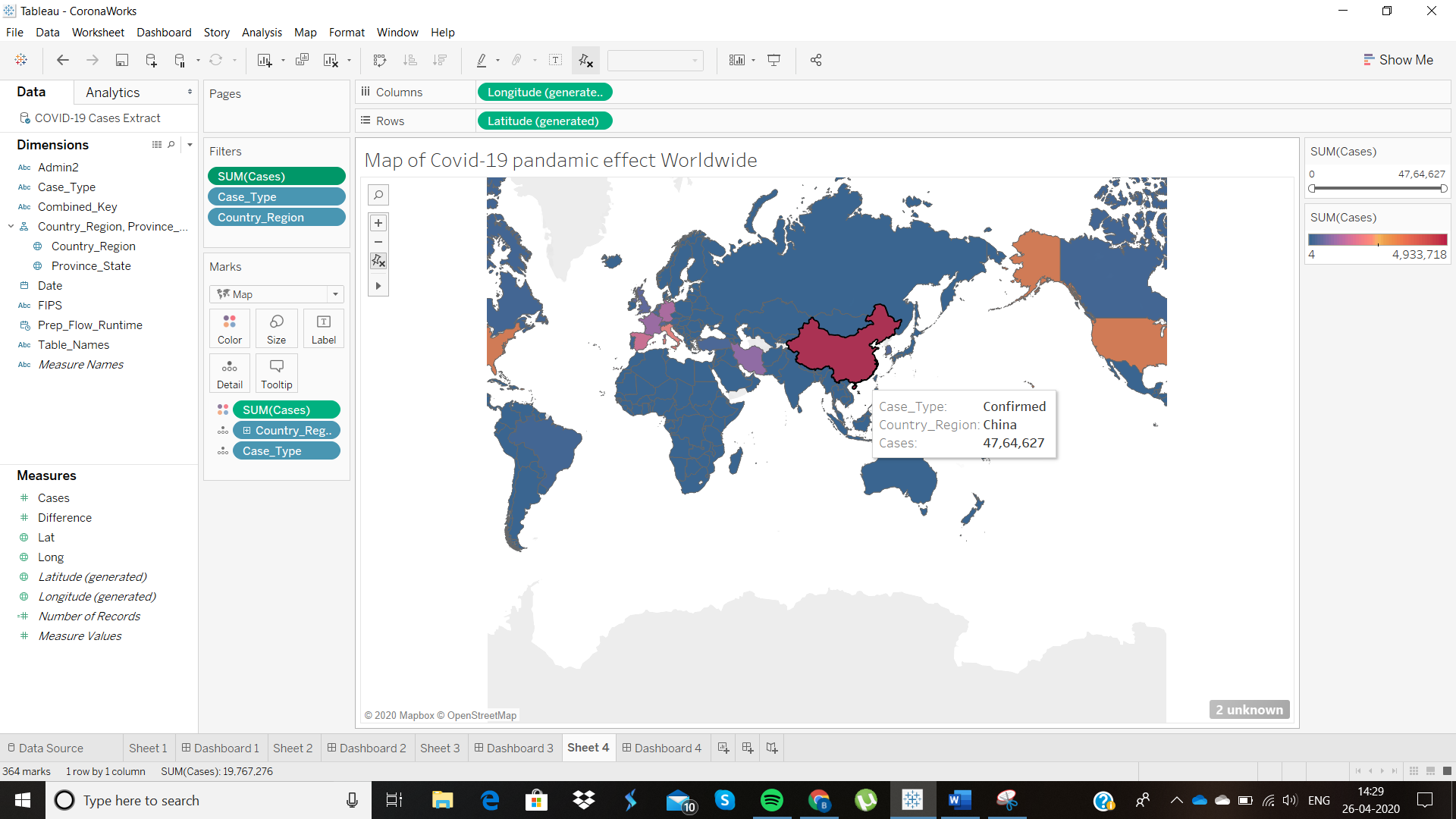
### Use one of the data visualizations features to represent the global pandemic with perceptual accuracy. State why the feature is relatively better.

Solution:

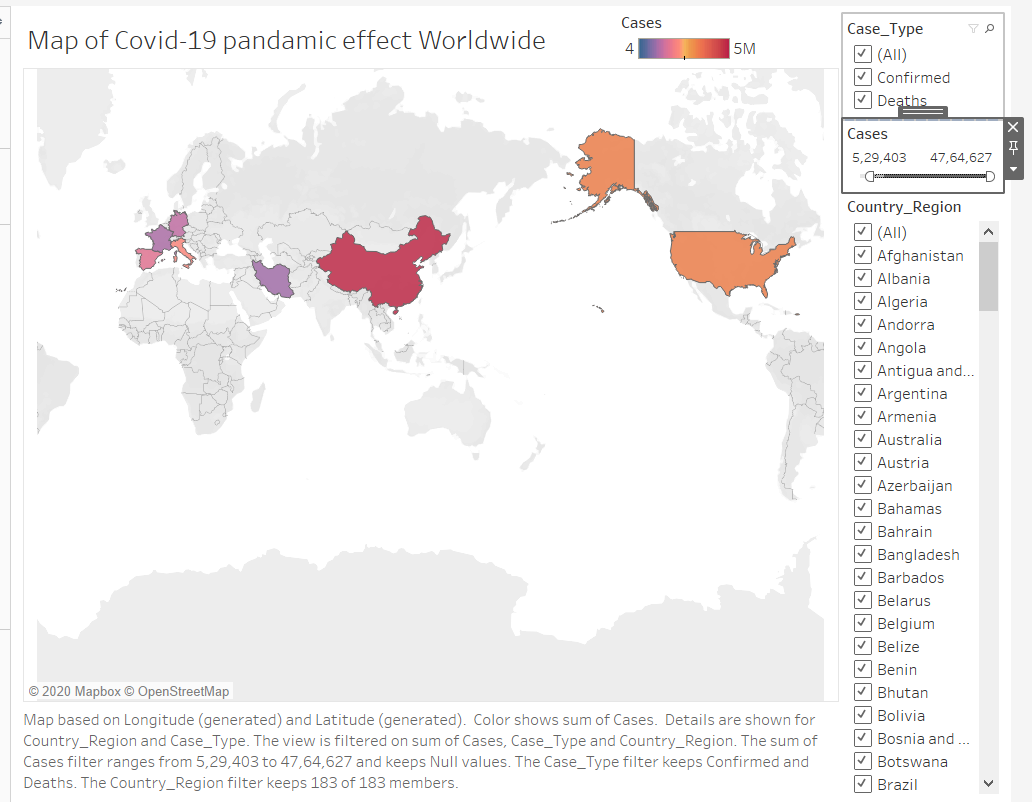
We use the Mapping feature to represent the effect of Covid-19 globally. Mapping adheres to the data visualization principles and adds to the perceptual accuracy.

We use latitudes and longitudes for mapping and along with filters like SUM(Cases) and Country\_Region. We also indicate the number of cases in different countries through colour mapping to show the difference and a scale is provided for comparison.



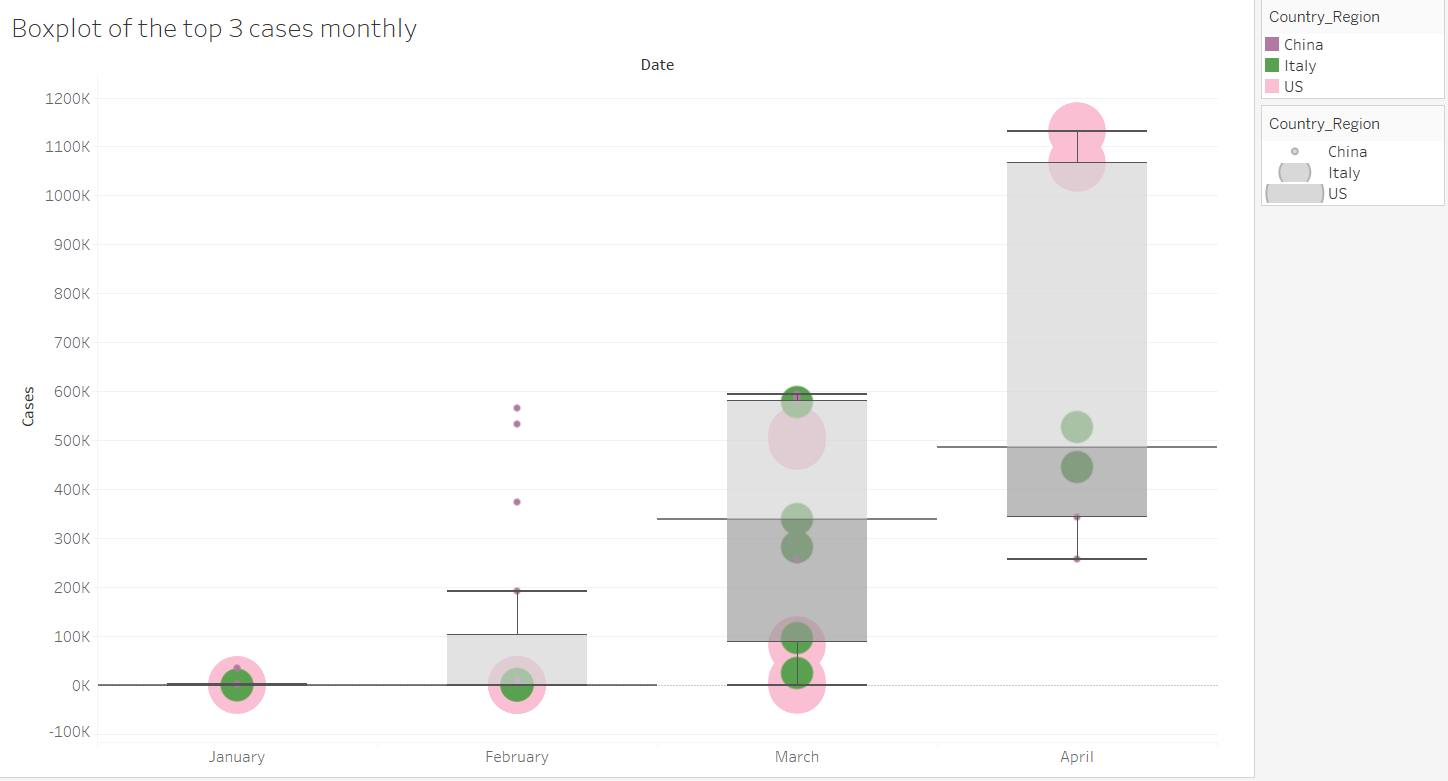
Colour mapping indicates the magnitude to which a country is affected by the Covid-19 virus strain. In the above example, we see the number of confirmed cases in Chine is close to 5 million as of April 5th. This magnitude is represented in red, and we can see from the colour meter that red indicates a high number or a problematic value. Let us look at the dashboard for the overall mapping and the summary of the mapping.



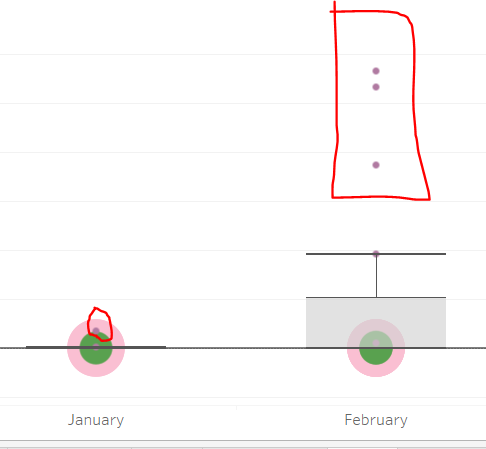
### Is the data influenced by the presence of outliers and if so, how far is it influenced? Take an example of top 3 countries with highest cases and prove the same.

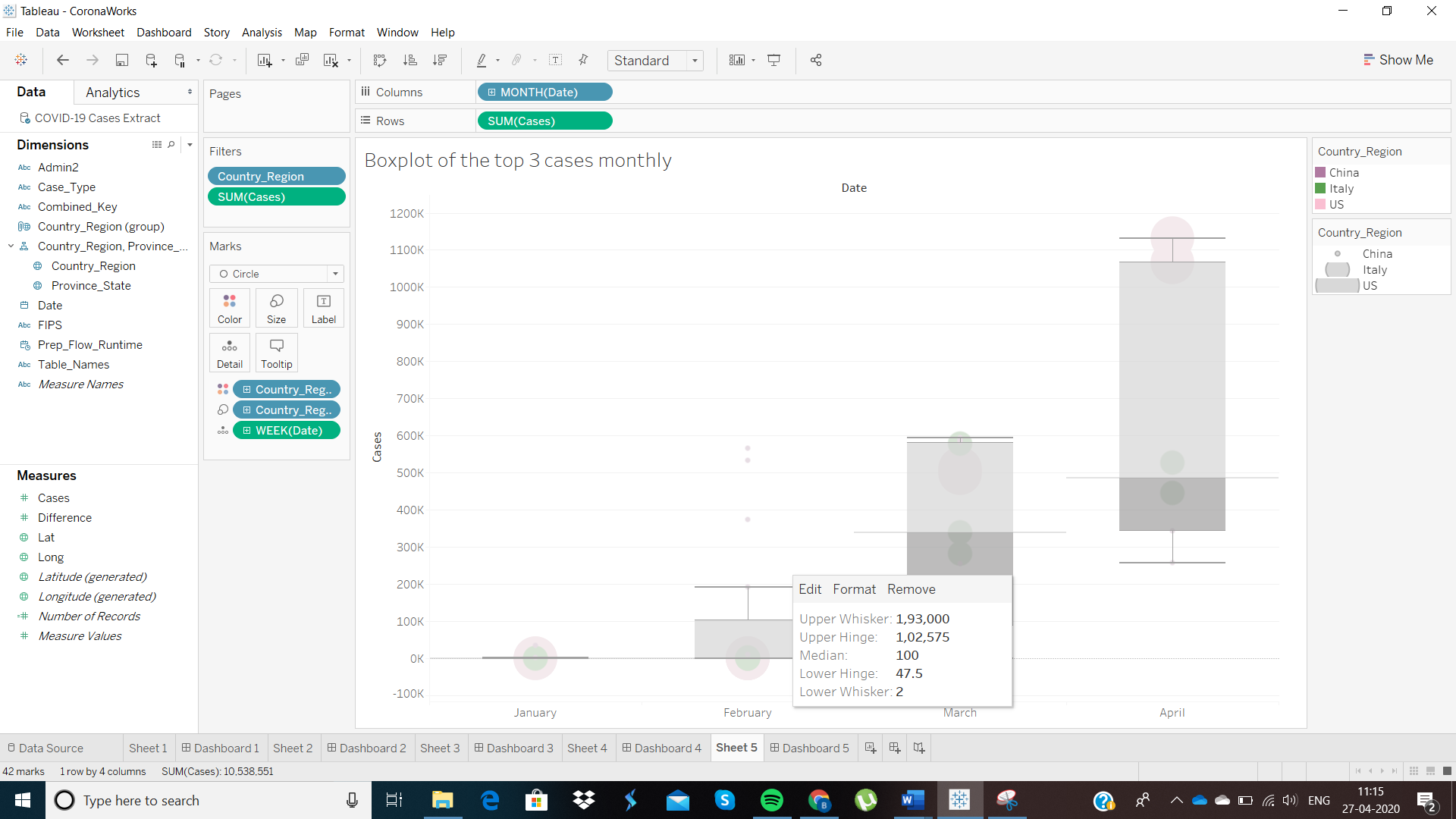
Solution:

We use boxplot to identify the presence of outliers. To obtain a clear indication of the results, we plot the boxplot for four months, ranging from January to April, and project it weekly.

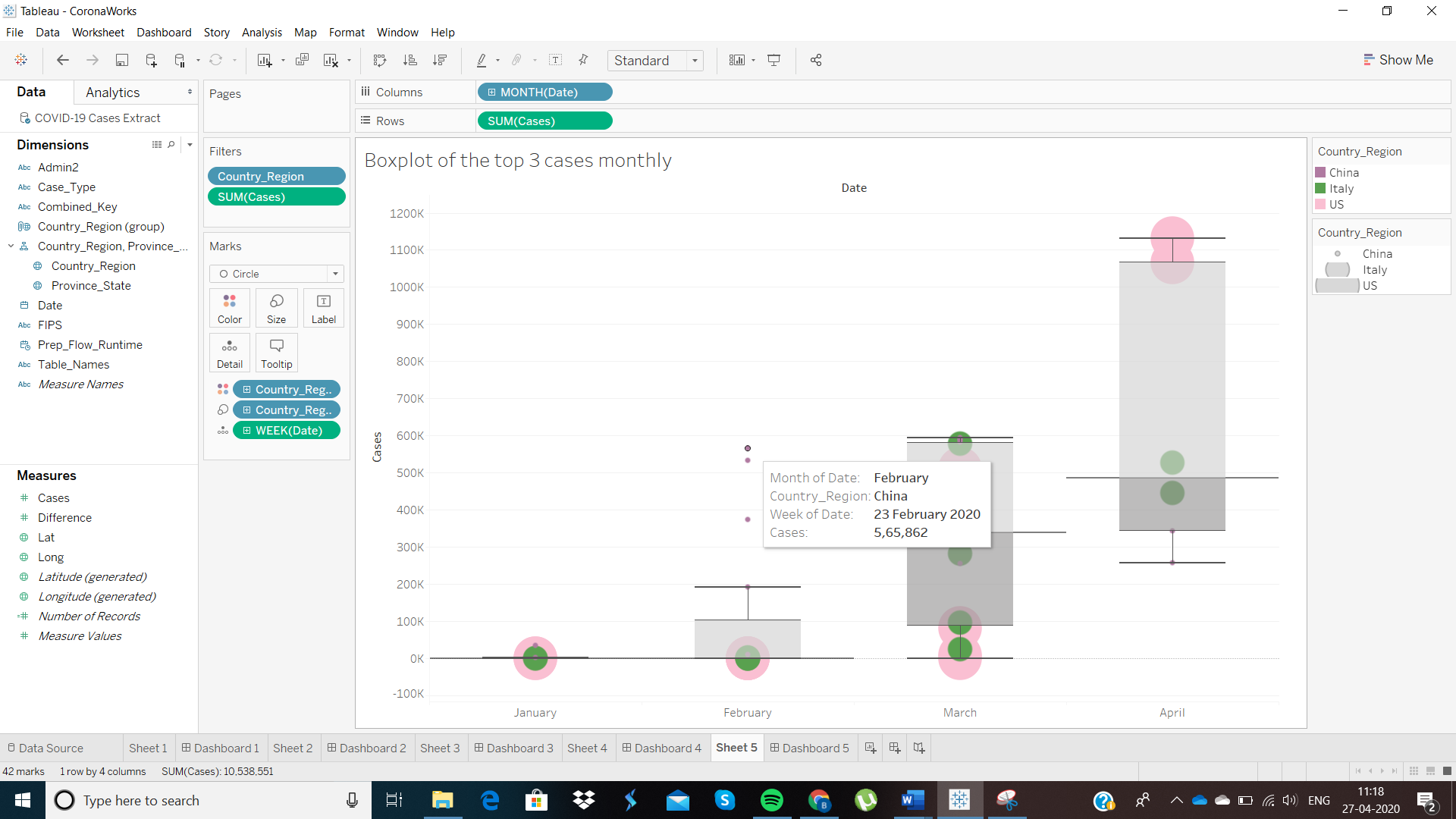


We see the in the month of January and February, we see the presence of outliers. The outlier effect is mostly because of the number of cases in China as it is relatively higher than that of US and Italy.





The Upper Whisker for the boxplot for the month of February is 193,000, the median is 100 and the Lower Whisker is 2. Now, let us look at the outlier value.



The highest outlier value for the month of February is 5,65,862, which is significantly higher than the Upper Whisker of corresponding boxplot. Therefore, the high number of cases for China, makes it an outlier from the statistical point of view and therefore increases the upper whisker. The actual number of cases was higher than the expected number of cases. The variation is the distribution of number of cases across the three countries over the respective time-period adds to the outlier values.

# References

*covid-19-coronavirus-data-resources#content-container*. (n.d.). Retrieved from Tableau: https://www.tableau.com/covid-19-coronavirus-data-resources#content-container

*understanding-boxplots-5e2df7bcbd51*. (n.d.). Retrieved from Towards Datascience: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51