**CS7DS4 / CSU44065 Data Visualization 2019-20**

** Assignment 3**

** Student Name:** *Divyanshu Marwah*

 **Student No**: *19304526*

 **Declaration**:

"I have read, and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found athttp://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism ‘Ready Steady Write’, located athttp://tcd-ie.libguides.com/plagiarism/ready-steady-write."

# Introduction

Visual representations help in augmenting human capabilities to easily study trends, patterns, and outliers in data, and consequently make data-driven decisions. The data used for this novel visualization project is the Coronavirus, COVID-19 dataset. This is one of the most trending datasets amidst the current pandemic outbreak. The objective of this visualization project is to create an interactive dashboard of Covid-19 world statistics and time-series visualization of the trend in the cumulative confirmed, recovered and death cases reported across the globe. This is an exploratory visualization, giving insights into different trends and patterns in data. Different views have been added in the dashboard to visualize and extract distinct statistics and analyse the impact with different frames of reference.

The dashboard contains a view showing the latest number of confirmed cases in an orthographic layout, tabular format, and a treemap visualizations. Each of these three plots serves different purposes and eases out the task of analysing the latest numbers. In another view of the dashboard, the trend is shown of the cumulative number confirmed, recovered and death cases from the day of inception of this virus to date. It is represented with an interactive stacked bar chart and an animated time series trend plot. A time-series visualization of the ten most affected countries by this pandemic is shown in the third view of the dashboard, to examine the changes in the number of cases and clearly distinguish the pattern of these changes in different countries with respect to time.

# Description

**Data and Data types used:** Four CSV data files have been used for this visualizations task. Three [1] of these files contain similar data with just change in metric recorded with dates. The columns are the name of countries(categorical data), name of states/provinces of some countries(categorical data), latitude and longitude of the countries(numerical data), date wise number of confirmed cases or recovered cases or the number of deaths(numerical and temporal data) – every date corresponds to a single column in CSV. All these columns of data are used in the visualization. Fourth CSV data file [2] contains three columns, out of which we are using two of them, they are, country name(categorical variable) and three letter country code(categorical variable). This is used for aliasing the country names and used to refer to countries in the plots.

**Tasks and Visual Encoding Channel**:In the dashboard, three different views for different sets of data and analysis have been implemented. The tasks and visual encoding of these views is explained in details in the following sub-sections.

* ***3D-Visualization and Tree Map of the Latest Data*** – This view shows three different plots, first is the orthographic view of the globe presenting the heat map of the confirmed cases. This view analyses the variation of the impact across different countries of the globe. This clearly depicts the magnitude of the cases in the countries and also study the impact of the disease in a geospatial dimension. The colours in the countries are filled with a gradient, with the most affected country filled with the darkest shade of red and least affected with the lightest shade. This assists in studying the intensity of cases and also visually distinguish the impact of disease across the countries. The plot can also be zoomed in/out, selecting a box/rough selection to focus on specific parts of the globe and highlight the impact in a particular region. The legend for this graph shows the gradient of colours along with the number of confirmed cases with markers set at every 100,000 records.

The second view is an interactive table, showing the latest number of confirmed cases, recovered cases and death cases for each country. It helps in analysing the impact in a tabular format with exact numbers. This table allows the user to search for a specific country/region and get the number of cases. The table can be sorted on any of the numerical columns and retrieve the most/least affected countries based on the confirmed cases/recovered cases/death cases. The number of records fetched in a single page can also be varied with a drop-box helping in getting the most/least affected N number of countries. The table also works as a lookup table for the country codes used in the treemap view.

The third plot in this view is a treemap. It represents the affected countries in a rectangular box in the grid format with each country represented by a rectangle in the grid. This represents the same data as in the other two plots but in a different view with a different purpose. This plot serves the purpose of scanning the impact on different countries in a single pane with a glance. The size of the rectangle in the grid is in proportion to the number of cases, that is the largest rectangle in the grid will depict the maximum number of cases and the smallest rectangle shows the least affected country. The colours in the grid are visualized with the gradient of colours changing between every position in the grid. The gradient ranges from light salmon pink to dark maroon shade. And this is also proportionate to the number of confirmed cases, that is, the most affected country is represented by dark maroon shade and next most affected country with a lighter shade and least affected country with the lightest shade. A similar colour pattern is used in this plot with the 3D plot to simply correlate both the graphs. The gradient is also depicted in the legend of the plot, which points out the scale of the intensity of the number of confirmed cases. Also, when hover/clicked on the specific rectangle in the plot, the legend pointer shows the impact on the gradient scale, with a step size of 250,000 records.

* ***World Trend Charts***– This view displays an interactive stacked bar chart of the cumulative cases to date. This graph depicts the cumulative increment of confirmed, recovered and death cases with respect to date. It is represented in a stack format with confirmed, recovered, and death cases represented by individual stacks and allocating a single bar for each day. Hover and click on the bar shows the respective cumulative number of cases. A particular section of the graph like selecting only certain number of days can be selected and the graph scales accordingly. That is if we select like 10 days in April, so this zooms in and covers the entire axes of the graph and scales the chart in the given dimension. We can also zoom in/out, box select/rough select a particular section of the graph and highlight the trend in it.

Another plot in this view is an animated view of the cumulative cases. This graph plots the trend line chart based on the time series of the cumulative number of confirmed/ recovered/ death cases. This helps in studying the increment or dip in the numbers with the date. With this, we can easily analyse the trends of increase in aggregated cases across the world. Another interesting part of this visualization is the dynamic size of the pointer leading the line trends. This pointer increases with the increase in the number of cases and it is directly proportional to the number of the case on a specific day. The axes in the graph mark the months and the title shows the exact date of the plotted number. Different colours are used to plot confirmed, recovered and death cases in this graph. The same colour scheme has been used in the stacked bar graph to easily relate the numbers shown in both the graphs.

* ***Time Series Plot of Most Affected Countries –*** For this plot, the ten most affected countries by this pandemic are fetched and time series line trend chart is plotted for the number of confirmed, recovered and death cases. It is plotted in a similar fashion as the world trend time series plot.With this visualization, we can easily differentiate and study the change in trends of cases in these ten countries with the period of time. In this plot, the aggregated number of confirmed cases are fetched for every day and a time series lag is created and plotted for each country. Similarly, the plots are created for recovered and death cases. Months are shown in marker and the specific date for the number is shown in the title. The leading pointer in these line charts is directly proportional to the respective number of cases. Different countries are represented by different colours and all three graphs are plotted in synchronization with each other. This helps in comparative analysis of the trends in most affected countries, both within themselves and with respect to other nations. For example, it can be seen that confirmed cases and death cases firstly increases in China but the curve flattens and respective recovered cases chart shows an increase. Similarly, steep increase in confirmed cases is seen in USA crossing all other nations while steep increase in death cases is seen in Italy, which is later crossed by USA in proportion to the number of cases. [[1]](#footnote-1)

# Implementation

The visualization is implemented completely in R, using RMarkdown [3] and HTML widgets [4]. A broad overview of the tasks involved in implementation are as follows:

* Out of the four datasets, three data files representing the metrics are in wide format, and these are converted into long format. After converting the data in long format, all three data frames are then combined to a single data frame. This data frame is then merged with the fourth data frame based on the name of countries. The pre-processed data after this step consists of name of country, states, code, latitude, longitude, confirmed, recovered and death cases.
* For orthographic globe plot, map projections and options are defined in one layer. This layer is added to the plotly object. Finally the last layer in this plot is added defining the axes, colour scheme, location markers and the pre-processed dataset.
* In interactive table, values are selected from the dataset, arranged in an ordered fashion, and dimensions and options are given using DT library.
* For treemap graph, highcharter library is used, by defining the layers with the data to be plotted, colour-scheme, colour axis, label and legend values.
* In stacked bar chart, data is first transformed to get the cumulative cases date-wise. After that, a ggplot bar chart is created by defining the bars, aesthetics, colour fill, scales and theme. This ggplot object is used in plotly to convert it into an interactive chart.
* For time series plots, the data is firstly transformed into time-series lagged version and converted into long format. Following this, a ggplot line graph is plotted by defining data, mapping, scales, colour scheme, themes, title, legend and then transition layer is added with gganimate defining date as transition parameter. Further, the plot is animated by defining the frames, fps, and pauses. [[2]](#footnote-2)

# Conclusion and Future Work

This visualization dashboard presented an exploratory visualization of the COVID-19 dataset. It is created with an aim to identify different trends, patterns and correlation in the confirmed, recovered and death cases in different countries to date. Different visual encoding channels such as colour gradients, geospatial positions, size, and motion have been used to achieve the visualization tasks in this project.

The current view of the dashboard represents individual interactive plots showing different views of dataset. This can be enhanced by interlinking these visualizations, for example, in the first view, if a country is selected from tabular format, it can be highlighted in the globe as well as the treemap view. Similarly in the second view, bar graph can also be generated with animation and in sync with the time series plot. Adding these features, will make the dashboard more interactive and enhance user experience.

# References

|  |  |
| --- | --- |
| [1] | J. H. University, “COVID-19 Time Series Dataset,” Johns Hopkins CSSE, 2020. [Online]. Available: https://github.com/CSSEGISandData/COVID-19/tree/master/csse\_covid\_19\_data. |
| [2] | plotly, “Country GDP data,” 2014. [Online]. Available: https://github.com/plotly/datasets/blob/master/2014\_world\_gdp\_with\_codes.csv. |
| [3] | J. J. A. G. G. Yihui Xie, R Markdown: The Definitive Guide, Chapman & Hall/CRC, 2020. |
| [4] | K. R. Ramnath Vaidyanathan, “htmlwidgets for R,” RStudio, Inc., 2015. [Online]. Available: https://www.htmlwidgets.org/index.html. |

1. Visualization available at <https://www.scss.tcd.ie/~marwahd/covid-19-3d-plot-ts-animation.html> [↑](#footnote-ref-1)
2. Code available at <https://github.com/divyanshumarwah/COVID-19-Time-Series-Viz-CS7DS4> [↑](#footnote-ref-2)