# **DATA MANAGEMENT DOCUMENT Data Team: Apurva Deore, Vishal Padma**

**I] Aspects of Data to be considered when choosing APIs**

Considering that we need the data that’d answer **statistical** and **theoretical** questions asked to the **Chatbot**; we further divided the data into two broad sections.

**DYNAMIC/LIVE DATA** 🡪 Statistics of COVID-19 cases [real time scenario, updates happen regularly]

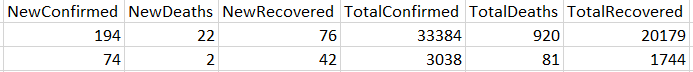
-Total cases

-Active cases

-Recovered cases

-Deceased cases

-Tests done



**STATIC/OLD DATA** 🡪 Data/facts about COVID-19 [stored-and-forgotten scenario, doesn’t need updates]

-Symptoms of COVID-19?

-Healthy habits to be followed?

-Whom to contact for testing related situations (some helpline numbers)

**II] Research conducted on the following APIs**

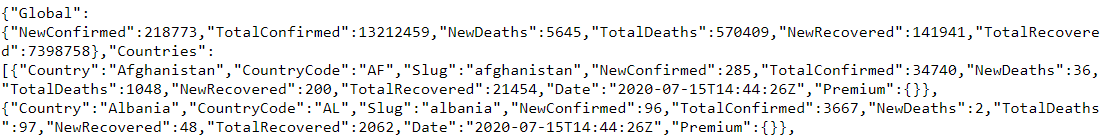
Initial research was done on various APIs such as below,

* Johns Hopkins COVID-19: ( <https://api.covid19api.com/summary> )
* European Centre for Disease Prevention and Control: COVID-19
* Ireland specific data (<https://data.gov.ie/>)
* Tweets from Twitter
* YATKO API (<https://api.quarantine.country/api/v1/summary/latest>)
* Bing-COVID-19 Data from GitHub (<https://raw.githubusercontent.com/microsoft/Bing-COVID-19-Data/master/data/Bing-COVID19-Data.csv>)

**III] DATA COLLECTION**

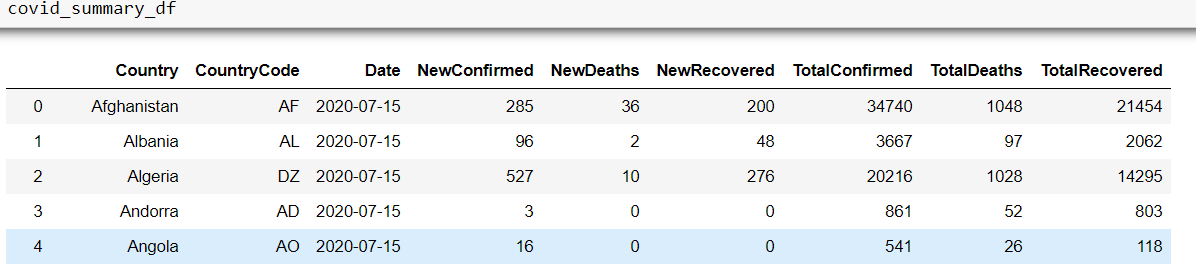
***API 1: Johns Hopkins***

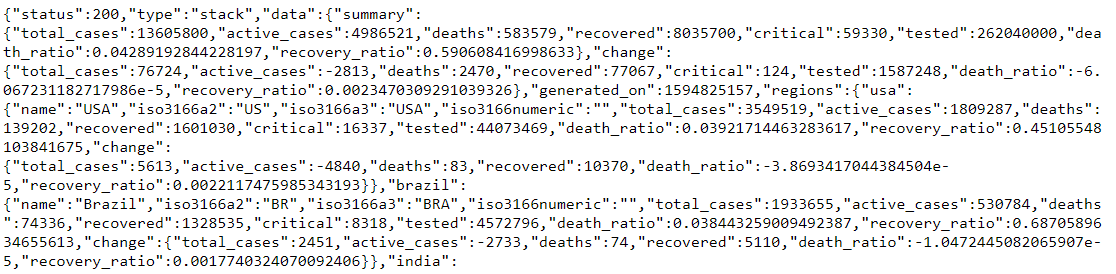
Gist of how the data looks to begin with:



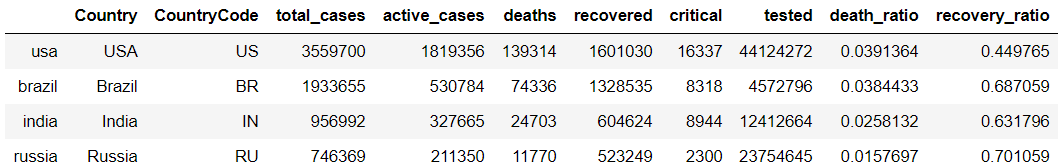
* It has all the data that is required for us to work on, ie. A lot of information about the statistics of MANY countries to start with.
* It gives the information specific to countries starting from the month of January until the latest updated date, ie. If we’re getting the data for the date 16th June, we’d get the data updated until 15th June. So basically, it provides us with real time data! This could be useful when we’re plotting visualizations for various countries (to show trends that the cases follow).
* We have already collected raw data pertaining to the total cases, recovered cases and deaths for about 186 countries as of now. We have also collected the time series data pertaining to the same 186 countries (for each day until today, starting from 22nd Jan, until the most recent update date of data collection).
* There were some data discrepancies, where the data was not recorded or recorded as wrong. To handle this we have replaced the statistics of the previous date wherever required and in some cases the active cases for each country were recorded falsely. So we calculated the active cases on a given date(where it was recorded falsely) by subtracting the death cases and recovered cases from confirmed cases.

Main data frame from the Johns Hopkins API:



***API 2: YATKO***Gist of how the data looks to begin with: 

* The data found in Yatko API is a summarized statistical data, where there is a total of 224 countries.
* There are several columns in this data and most of them are similar with the ones in Johns Hopkins API.
* The additional columns like death ratio, recovery ratio and tested cases for countries is also present in this data.



As of now we have two APIs which give summary for different countries but both have some different columns, so we decided to integrate both the datasets of summary and create a single dataset which will contain all the necessary data from both the APIs. So, now we have integrated the summary data for 186 countries from both the APIs and the final merged Dataset set has below mentioned columns for each country which will be later used in the visualization.

-Country Code

-Country

-Date

-New Confirmed

-New Deaths

-New Recovered

-Total Confirmed

-Total Deaths

-Total Recovered

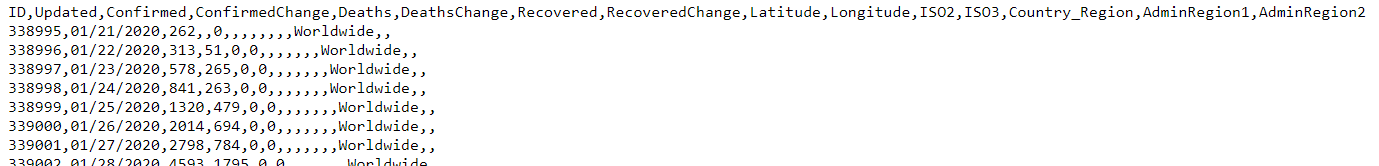
-Death Ratio

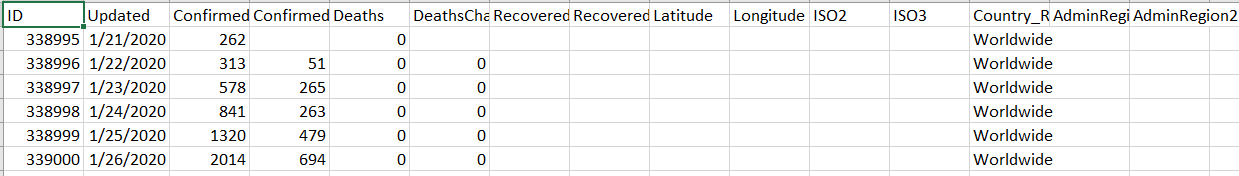
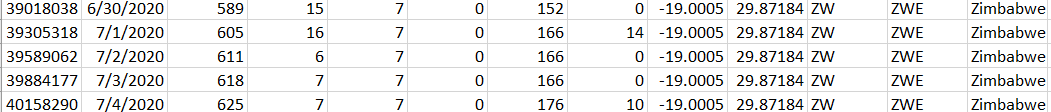
-Recovery Ratio

-Total Cases Tested

-Critical Cases

***API 3: BING***

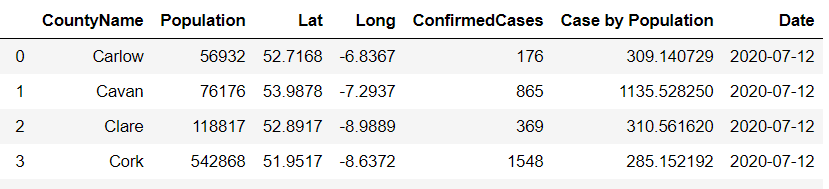
The 3rd data which we will be looking into is the data source from Bing COVID-19 wherein, the data is collected from many different sources and the same is converted to time series data. This data records Confirmed cases, deaths and recovered cases. Data is recorded when the first case is seen in that particular country. It also records the data for each date and for some countries the data is recorded for each date for each state/province. But some countries don't have data recorded according to the state/province. One interesting thing is that this API also records data for the whole world for each date. We have saved this data and plan to use it to handle any data discrepancies that could occur in the near future in the Johns Hopkins data. So in such a data-related mishap, we can refer to the Bing data and fill the missing or bad data into Johns Hopkins from this data to maintain and fill the gap in the datasets which will be used by the Chat-Bot or the world map.

***Dataframe:***

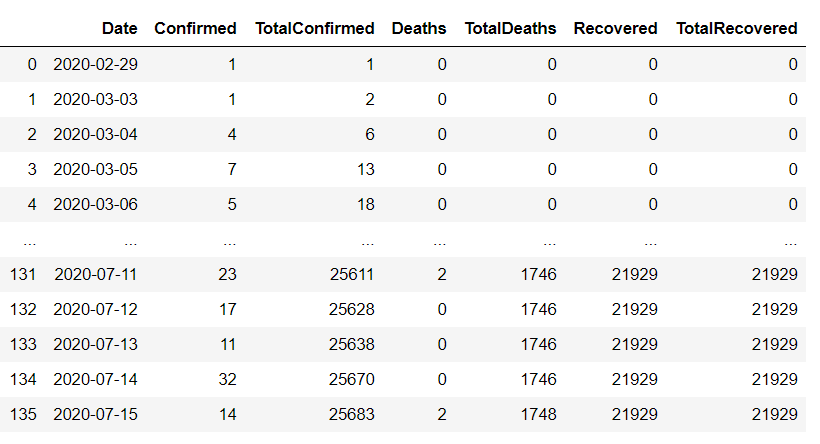
***API 4: Ireland Data provided by data.gov.ie***

This data source provides us with the data focused on Ireland only. It provides the statistical data for Ireland’s 26 counties separately as well as for the country as a whole.

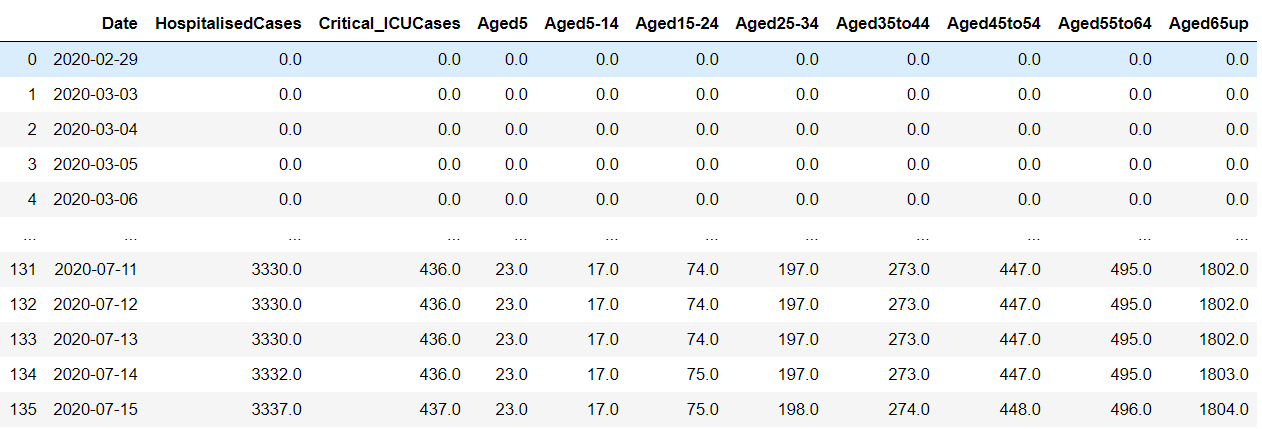
**Dataframe1**: It consists of the country’s population data (pertaining to each county), **confirmed** cases of COVID-19 till date and **Cases by population data.**



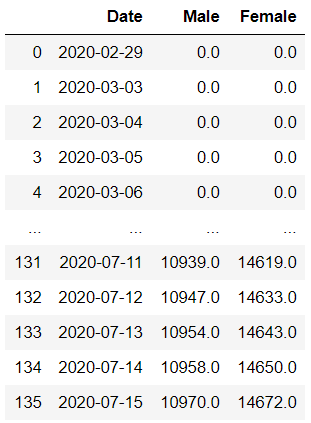
**Dataframe2:** This data frame consists of statistical data revolving around the **main types of cases.**



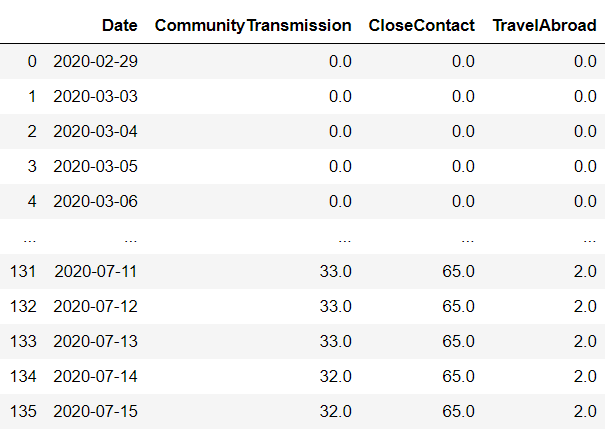
**Dataframe3:** This data frame consists of information regarding **Hospitalized Cases**.

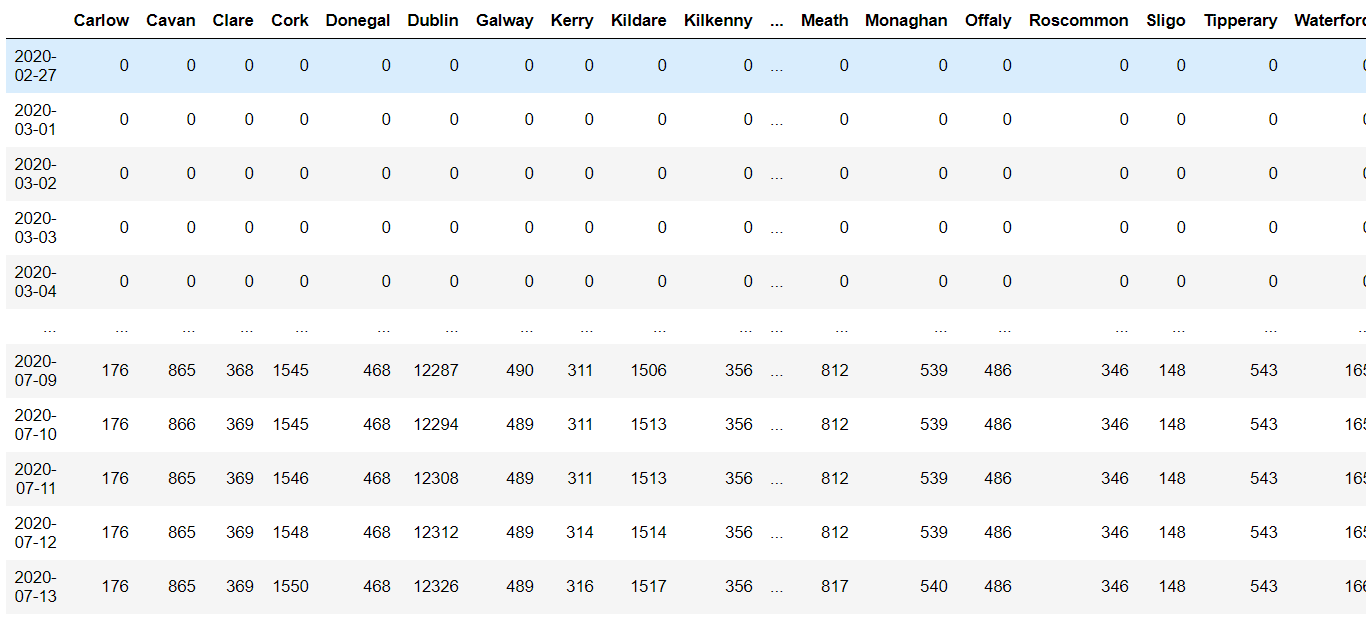


**Dataframe4**: This data frame consists of **Gender-wise** COVID-19 cases differentiation.



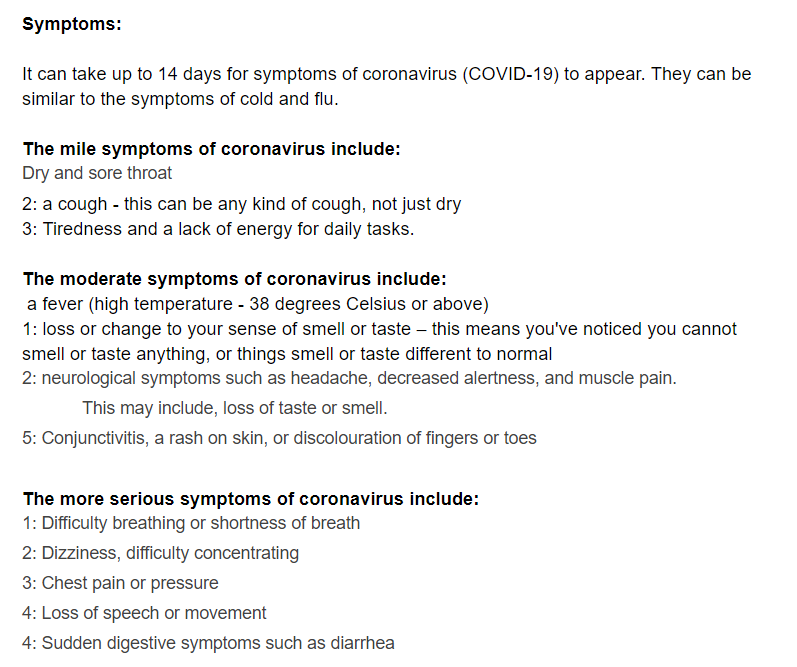
**Dataframe5**: This data frame consists of the **reason of** **spread categories**.



**Dataframe6**:This data frame consist of the **County-wise Confirmed Case related statistics**. 

Apart from this real time data, we look forward to collecting static data (which requires no updation) pertaining to the symptoms of COVID, healthy habits and so on.

Just a gist:



**IV] Data Cleaning/Pre-processing**

1. **Johns Hopkinns**

Pre-processing:

1. We need the summary data to be clean enough in the matter of Timestamp and other elements.
2. We have dropped only one column from the summary data which is “slug”. This column was used to refer country names.
3. Next is the country data which we collected. This data has many columns which were null so they were dropped.
4. The next thing was to convert the data into time series data.

=> **Now we have two main datasets pertaining to the Johnss Hopkinns API**

* One with the global summarized statistical data for 186 countries.
* One with the time-series data with information starting from the last week of January until the latest date of update.

For the cleaning/preprocessing we planned and did the following,

Summarized

* To clean this data frame, we removed the unwanted columns like slug, it wasn’t really required, re-arranged the columns to make a better looking data frame, changed/modified the date column such that it represent only the DATE and not the timestamp).
* Created a separate column to store the Country and another customized column that'd store the Active, Confirmed, Recovered and Deaths for every country.

So for every country A, there would be 4 rows dedicated to the Country A, e.g. Afghanistan

eg, Country Afghanistan : Confirmed  
       Country Afghanistan : Active  
       Country Afghanistan : Deaths  
       Country Afghanistan : Recovered

* So when we'd be referring to the cases in particular, we'd use two keys (Country name with the cases column) to identify the same.
* We also added additional columns to separate the date into the respective DAY of the week and MONTH of the year. (Didn't consider the month, it's constant: 2020). Used the Calendar package for the same. Additional columns added separately for each country's records.

**B. Yatko API**

i) The Yatko API provides with the columns:

**name, iso3166a2, iso3166a3, iso3166numeric, total\_cases, active\_cases, deaths, recovered, critical, Tested, Death\_Ratio, Recovery\_Ratio, change**

After considering the Yatko API has additional data as compared to the Johnss Hopkinns data, i.e. **critical, Tested, Death\_Ratio, Recovery\_Ratio, change** in particular, so we decided to combine those two datasets and keep the necessary columns integrated in one dataframe.

**ii) Merged Dataframe: Yatko + Johnss Hopkinns API**

This is the merged dataframe in which we performed pre-processing and cleaning in order to make the Yatko dataframe compliant with the main dataframe (Johns Hopkinns).

In the pre-processing phase, we took care of the countries like Namibia (countryCode “NA” was getting counted as a null value when reading the dataset), differences between namings between the two APIs (USA/United States of America ; UK = United Kingdom), other countries like Congo had two different names assigned for same values, CAR = Central African Republic. To make the dataframe look more appealing and easier to interpret, we renamed columns and cleaned it further.

Now the merged dataframe consists of the following columns giving information for 186 countries:  
**CountryCode, Country, Date, NewConfirmed, NewDeaths, NewRecovered, TotalConfirmed, TotalDeaths, TotalRecovered, DeathRatio, RecoveryRatio, Total Tests Done.**For a clear and concise view of the data sources, we’ve put up a simple flow diagram to explain the main sources and what all data records/columns from them we would be using going further.

**C. Bing API**

The Bing csv provides us with the columns similar to those provided by Johnss Hopkinns API. We’re considering the BING dataset so as to help us to handle emerging/probable data discrepancies (if any).

**D. Ireland Data**

Cleaning and processing done in order to get the data frames separated as and how required by us. Pictures of the data frame have been provided in the earlier section.

**V] VISUALIZATIONS**

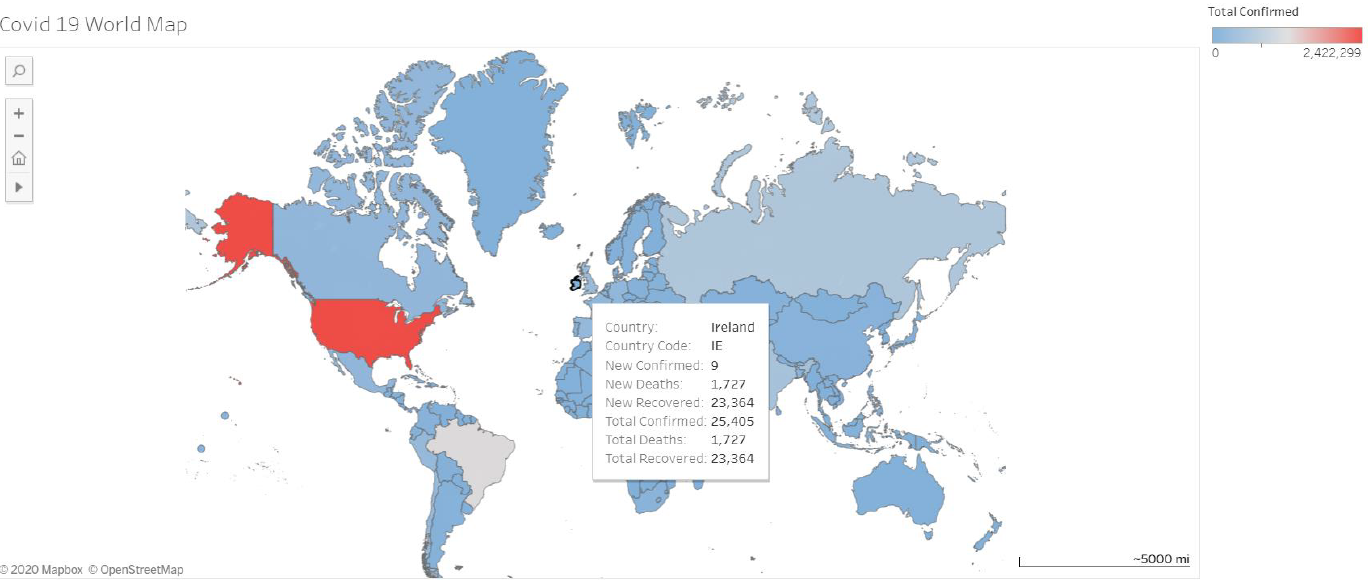
We’re considering **Tableau, Flourish and PowerBI** as the primary tools for creating visualization dashboards which would be embedded into our web-pages. We want to focus on providing useful insights revolving around the following (not limited to):

* Which countries have the most number of cases?
* Which countries have the best recovery rate?
* Which countries recorded a lot of deaths?
* What is the number of tests performed by each country?
* How many critical cases are present in X country?

1. ***TABLEAU***

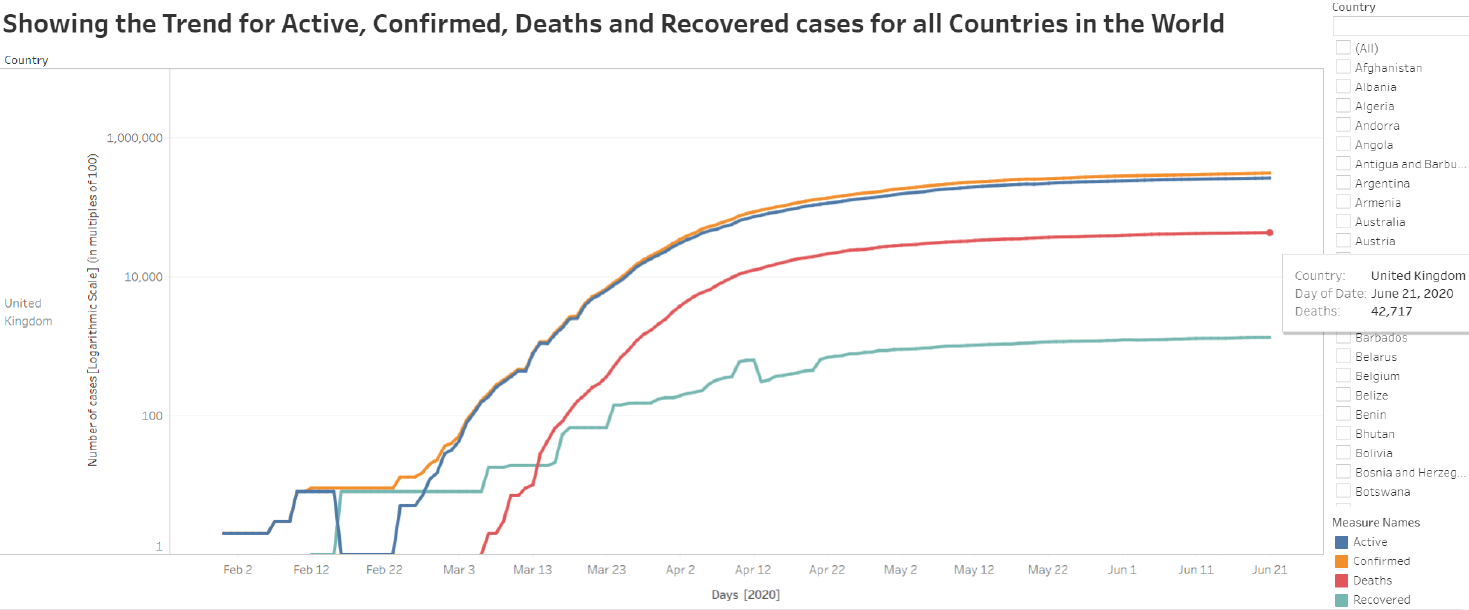
We’ve decided to make the following main types of visualizations

1) **Map related visualizations** (when a user hovers over a particular country, the map would show the user statistics for that country (ie. Details about active, deceased, recovered and confirmed cases). This would be displayed on the main webpage and uses the Summarized data which has info about cases for all countries affected by COVID-19.



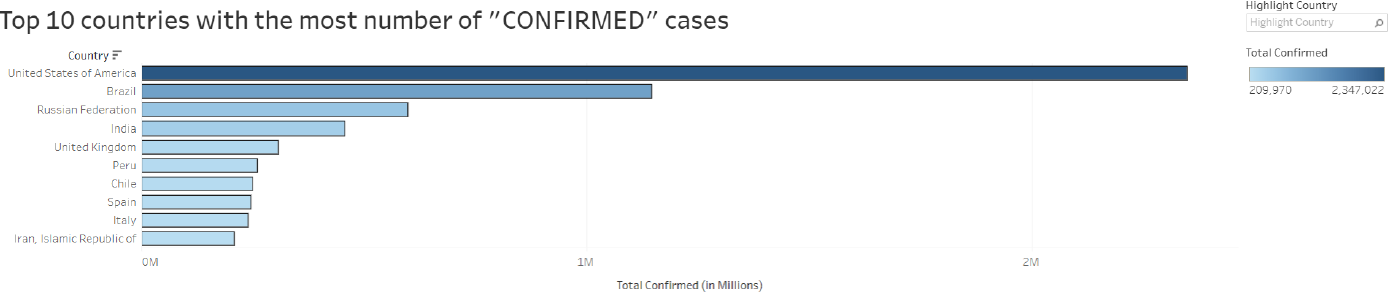
2) **Trend graph** : A visualization that’d show users the trend followed by a particular country. Basically 4 trend lines in one graph. (representing active, deceased, recovered and confirmed cases trend-lines).

* One worldwide (trends for the world’s total cases)
* One country-wise (filter on the country)



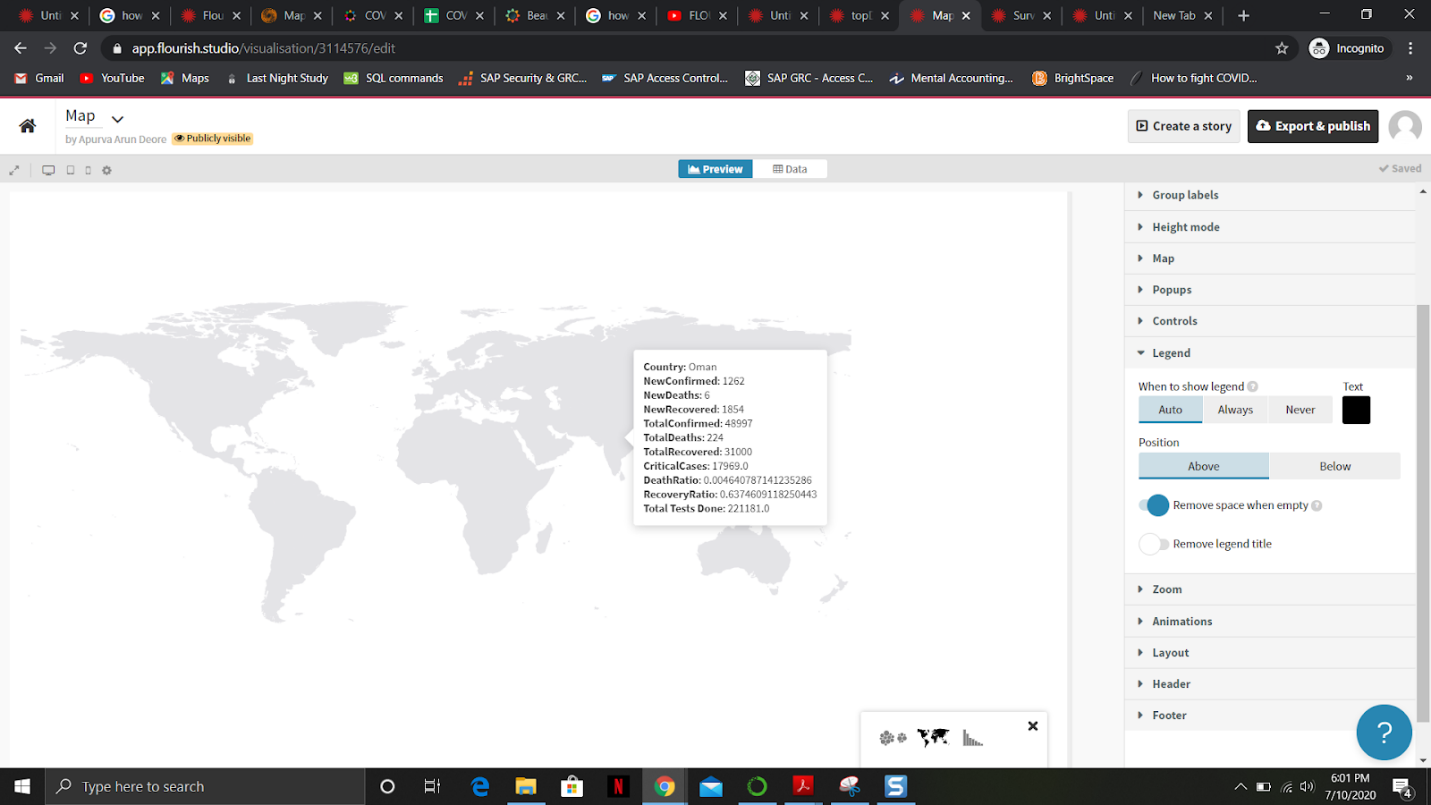
3) To show top 10 countries with respect to the cases as for recovered, confirmed and deceased cases. (considering the world-wide summarized cases dataset)

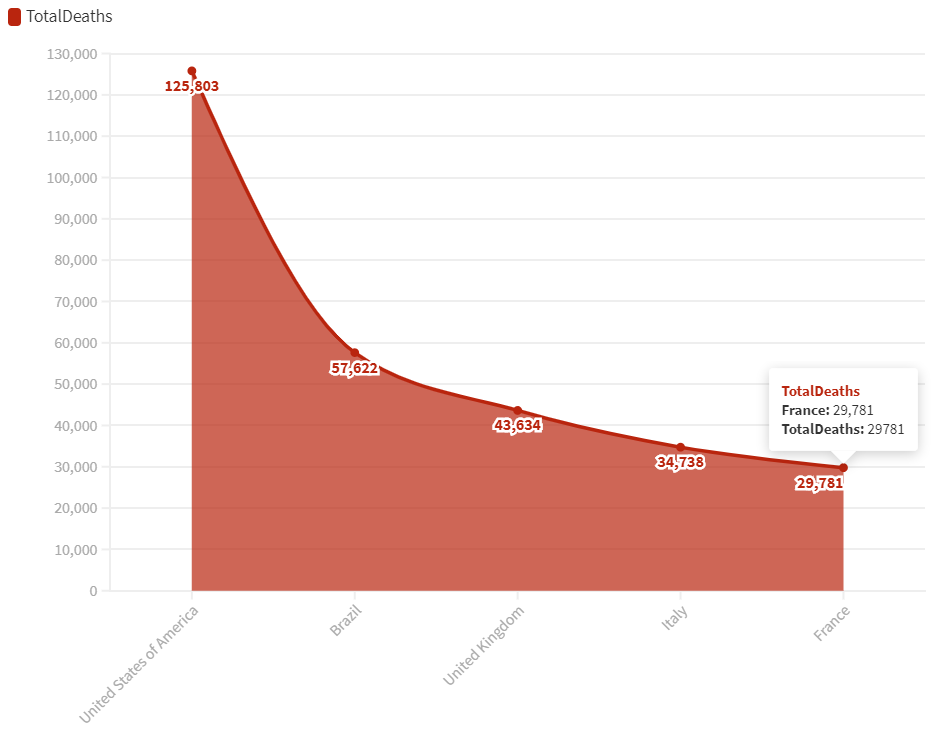
* Countries with rank in the Top 10 for Deaths, Recovered, Confirmed cases



1. ***Flourish***

Flourish is a tool that allows to create visualization dashboards/storytelling scenarios in a more fancy way. It provides certain map visualizations, to name some: survey maps, projection maps. We tried to create a map visualization similar to that done in Tableau to sort of compare which one is better. Some visualization we tried were as follows:

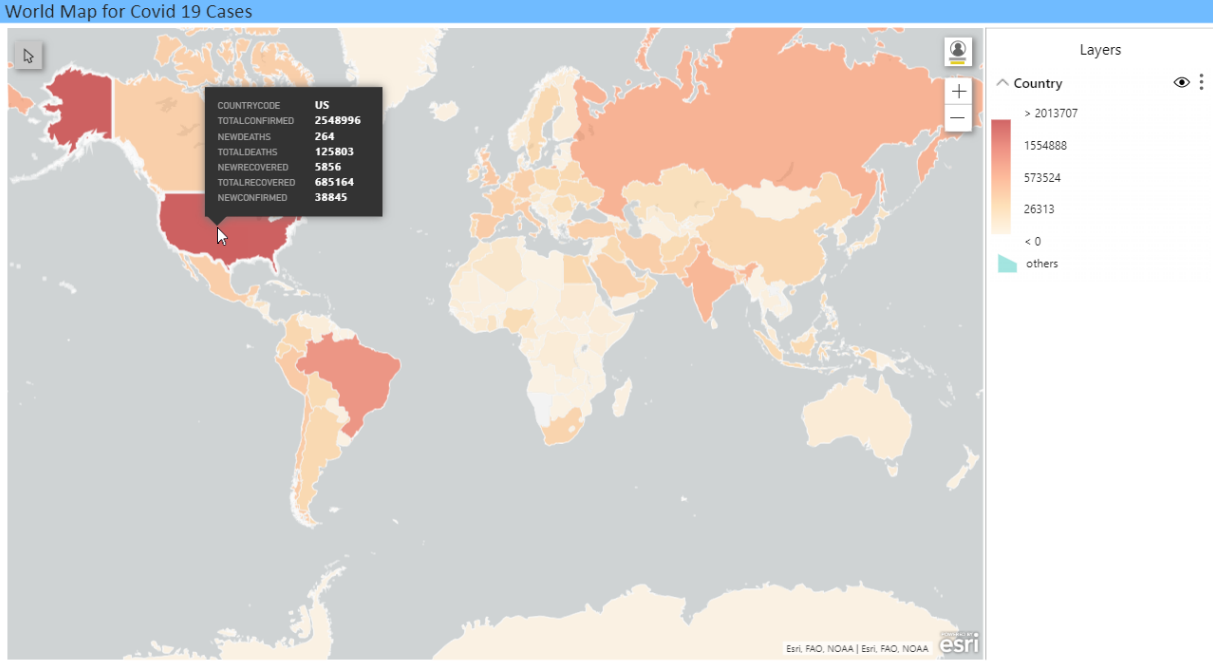


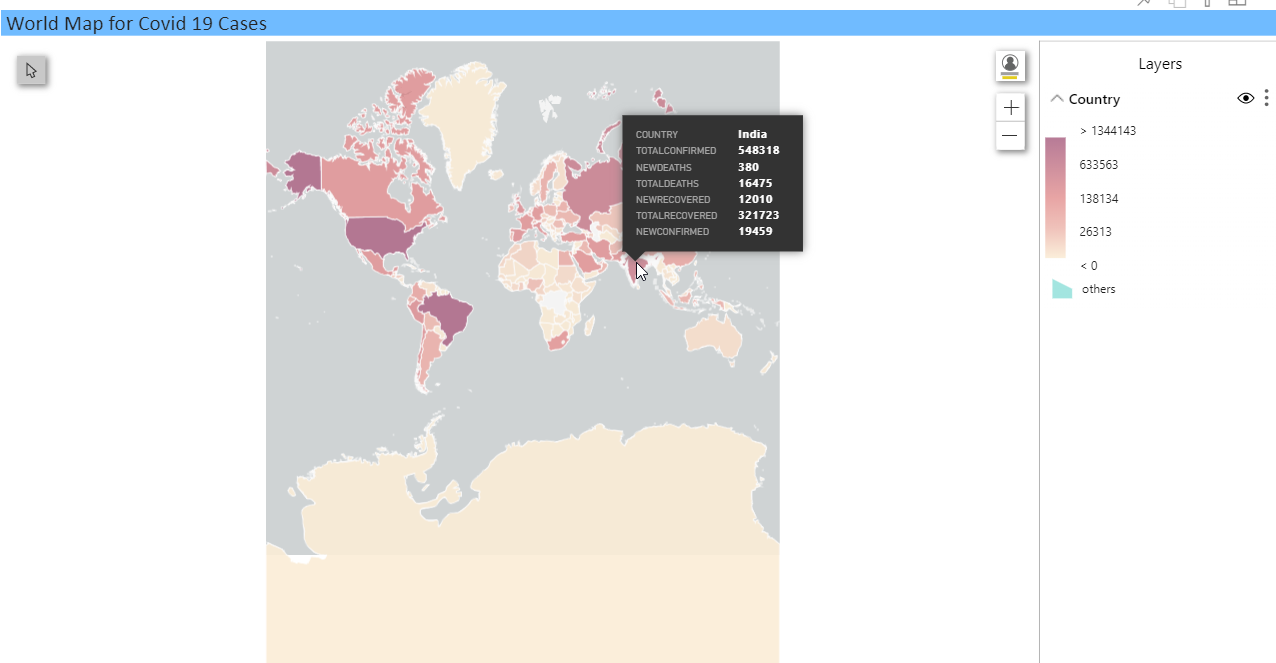


It provides a plain and bland view of the map with the statistical data displayed accurately, but it takes in a lot of unnecessary parameters for the visualization to be created and follows a very tedious approach. When dealing with a map, it asks for values like the geometry of a country, latitude/longitude values which makes it complicated unnecessarily.  Flourish could be used as a tool by those who write blogs, maintain fancy websites or someone who likes to create fancy stories from data.

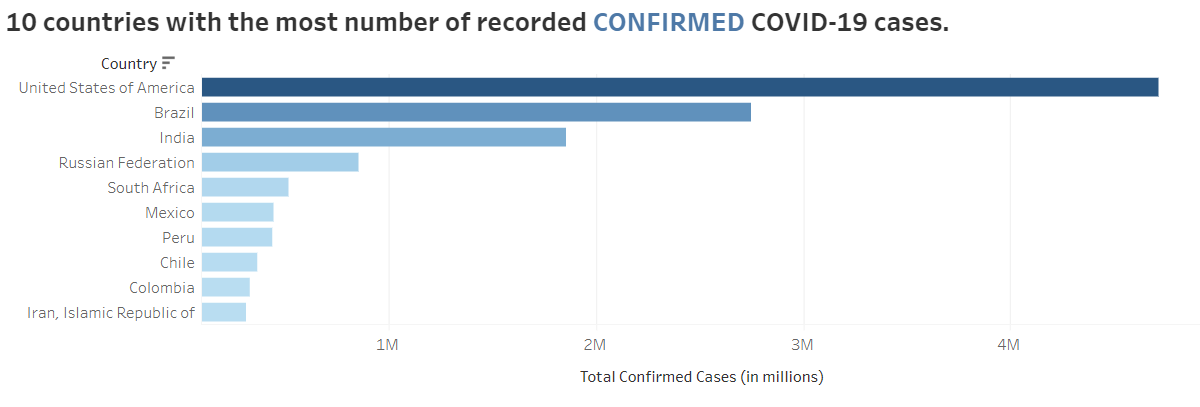
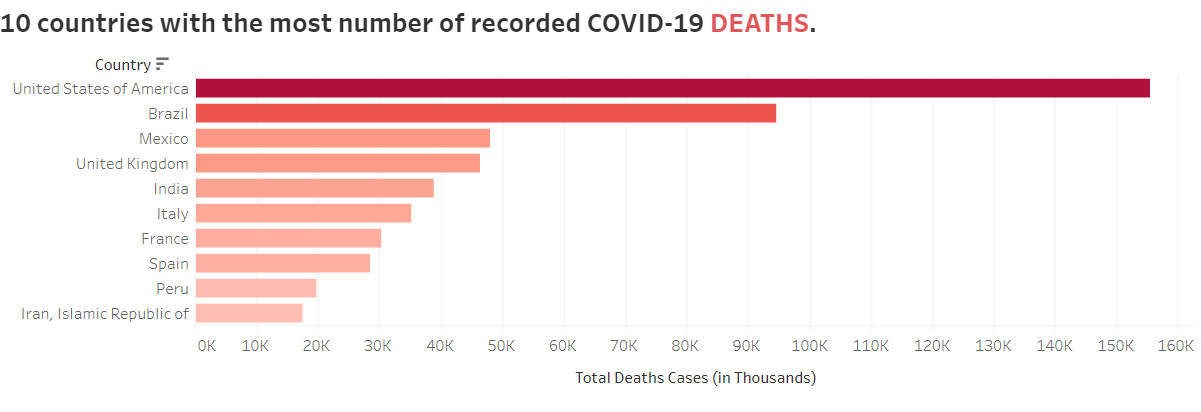
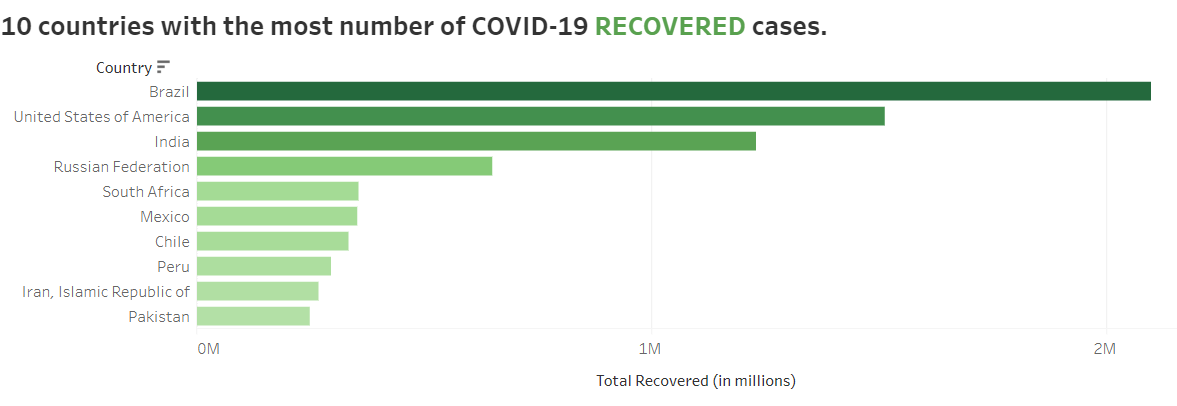
1. ***PowerBI***

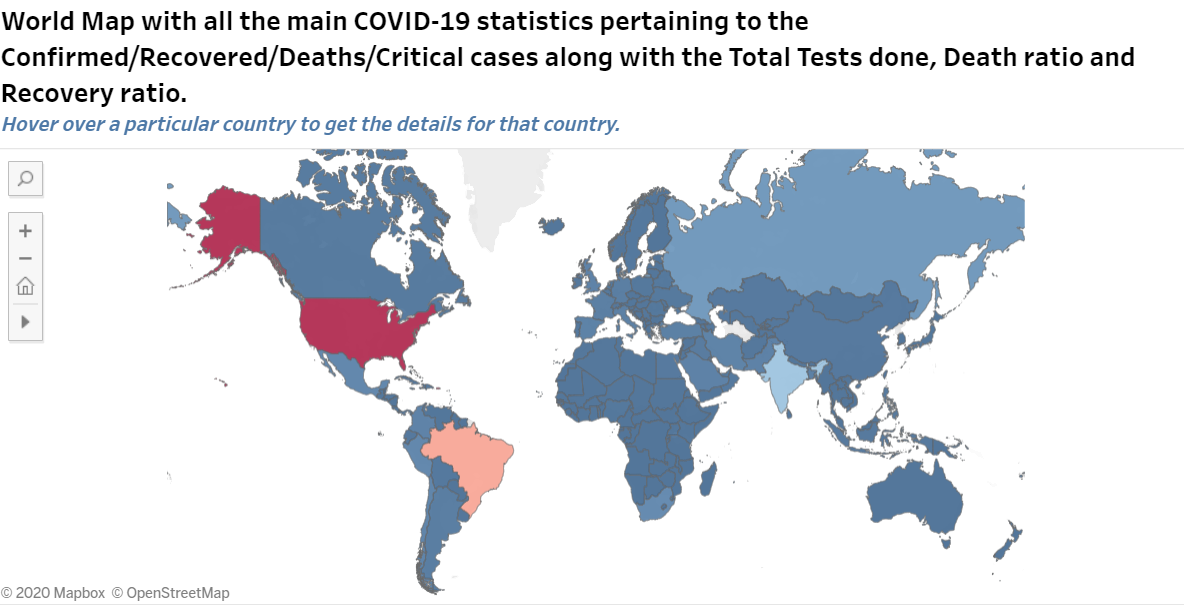
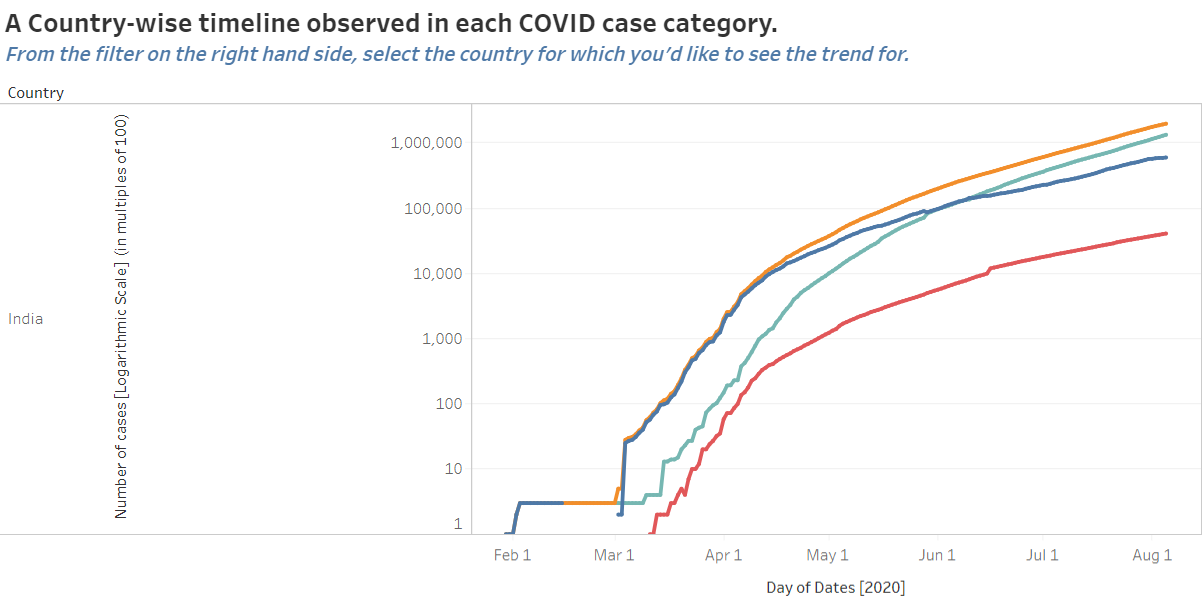
Tried the map visualizations with PowerBI and below are some examples of the same. PowerBI is an equally powerful tool as Tableau when it comes to delivering visualizations in terms of bar graphs, maps, pie charts, trends. It helps to create visualizations similar to that of Tableau.

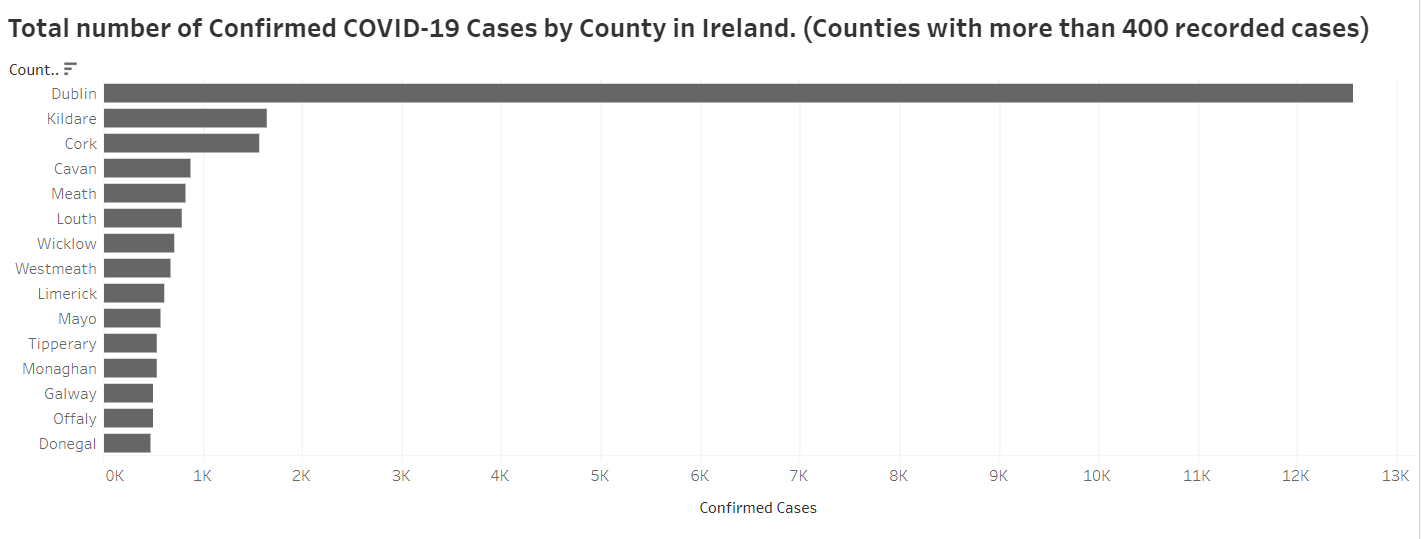
******

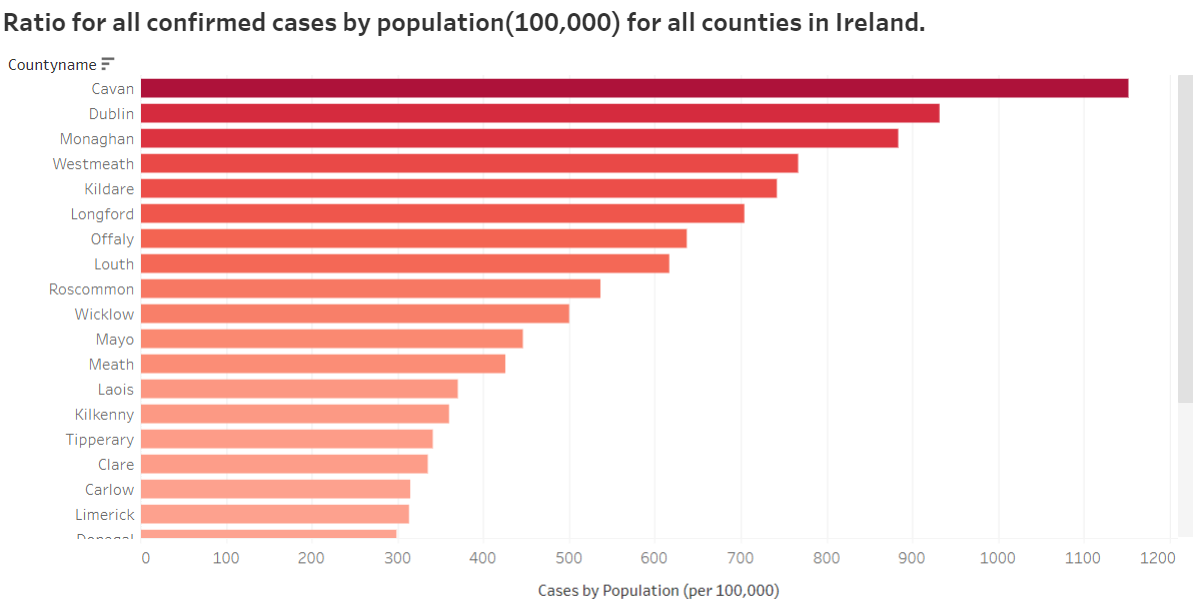


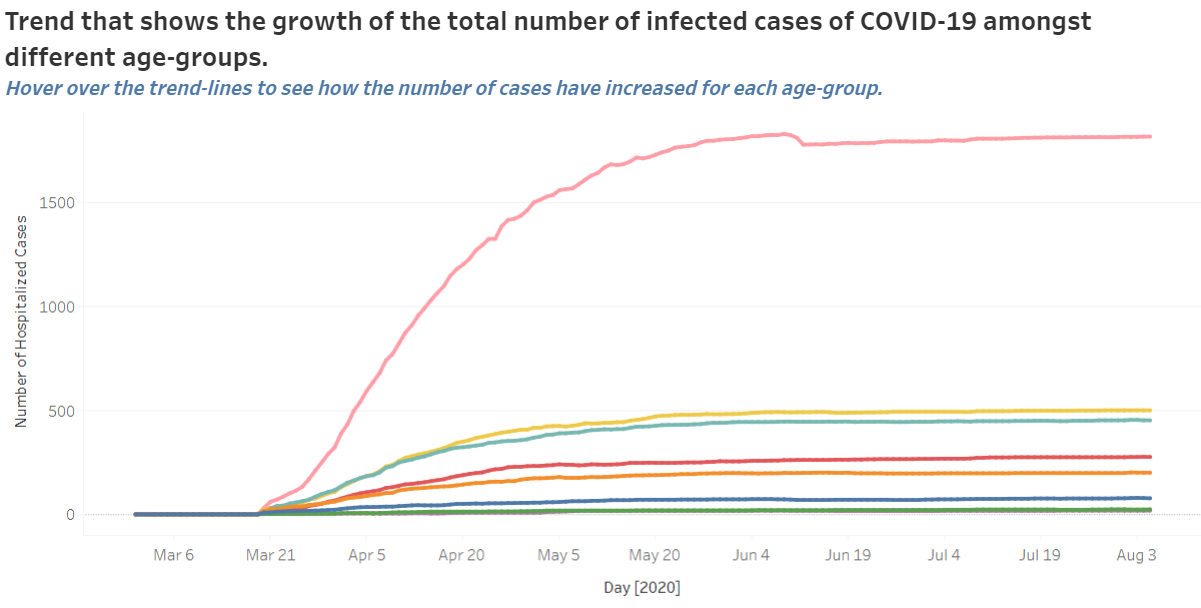
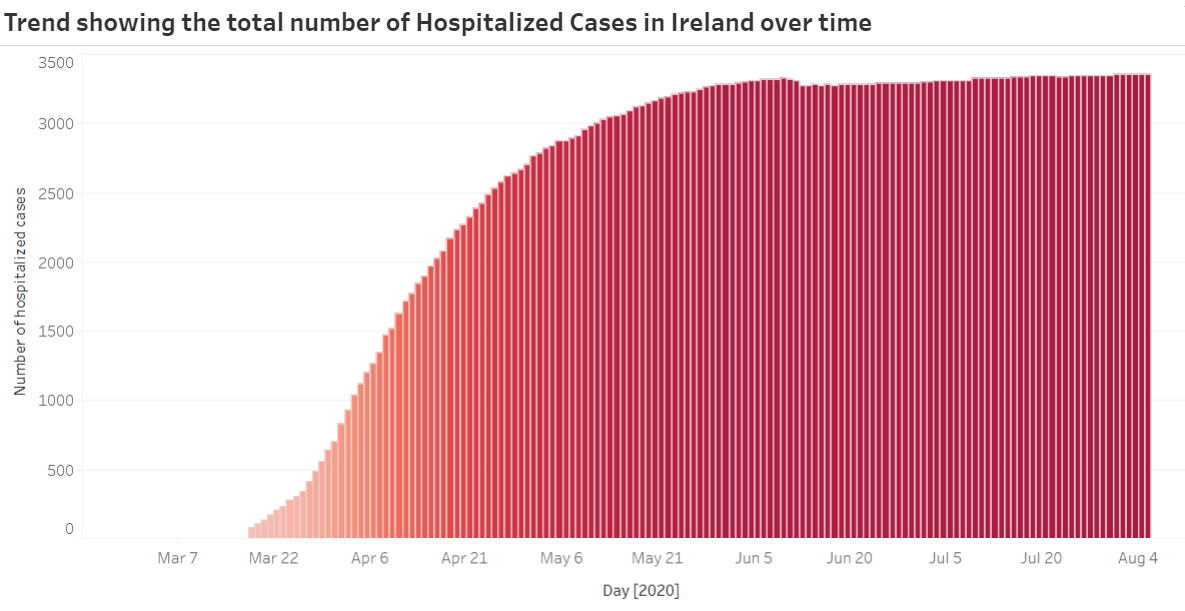
Conclusion: After having done a trial and error application of each of the above three visualization tools, we’ve decided to go ahead with Tableau as our main visualization tool because of the ease and simplicity it provides when trying to create a visualization. It takes in just the right amount of columns required to create a visualization without having the need to use data columns which aren’t required at all. Also, after having embedded the maps created using Tableau and PowerBI into our web-pages, it was evident that the one’s created using Tableau provide a clear and concise view of the maps and also look better than the ones made using PowerBI. So for now, we’re sticking with Tableau as the main tool for visualization whereas we’re keeping PowerBI and Flourish as backup options, just in case we come up with some additional visualizations which can be made better-looking with PowerBI and Flourish.

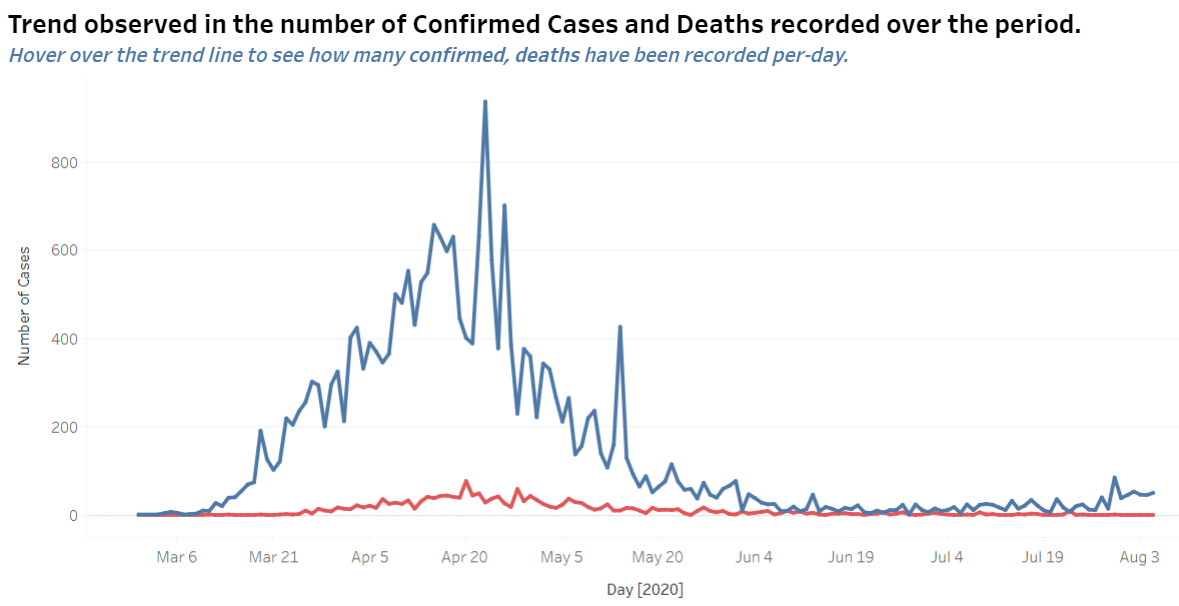
All our final Visualizations so far!

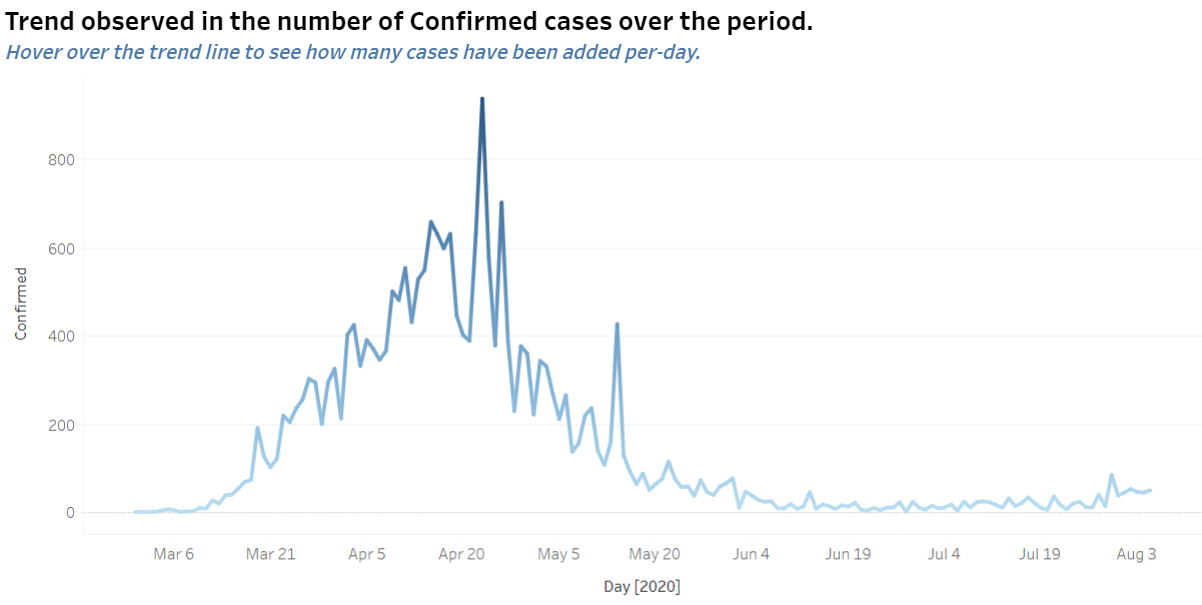


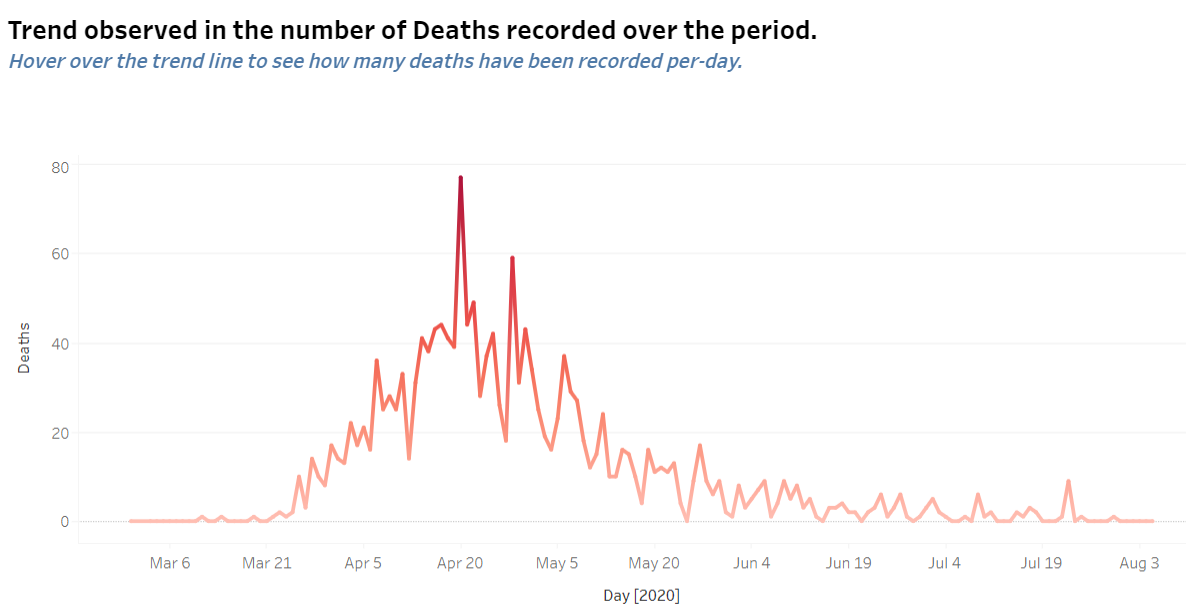


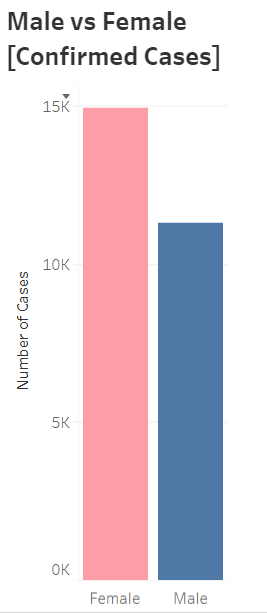




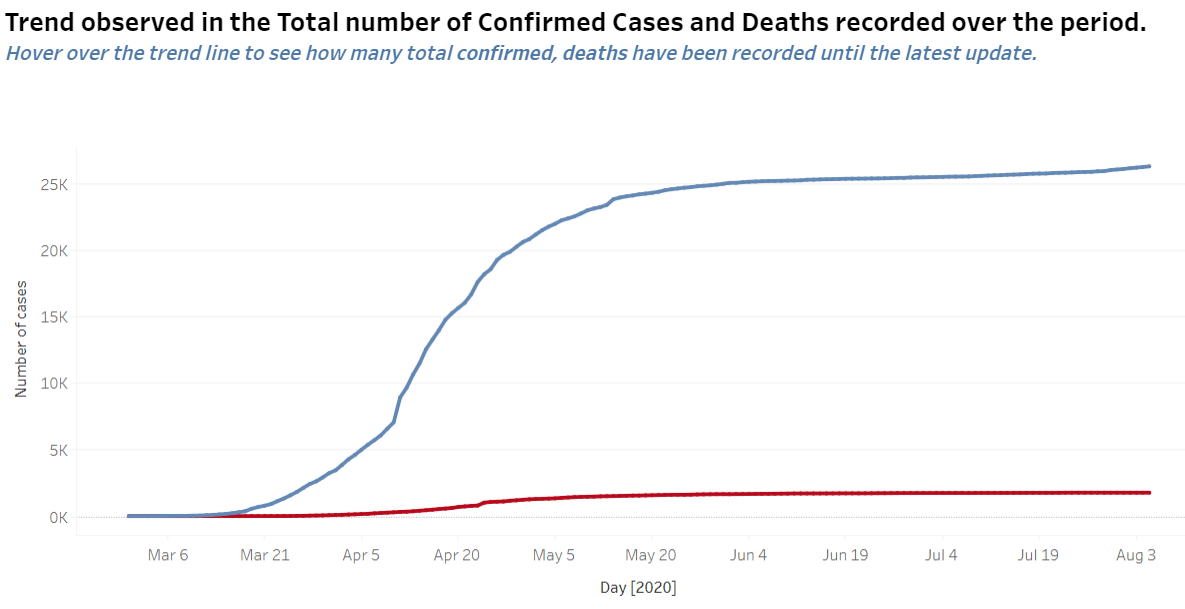


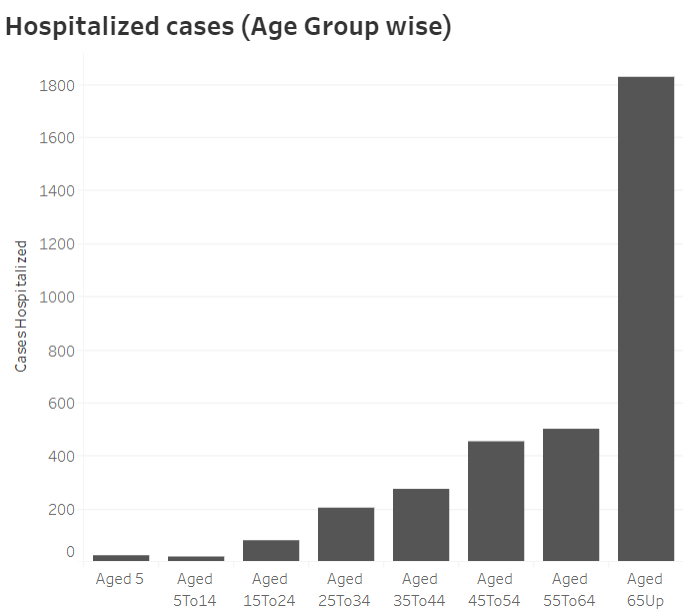


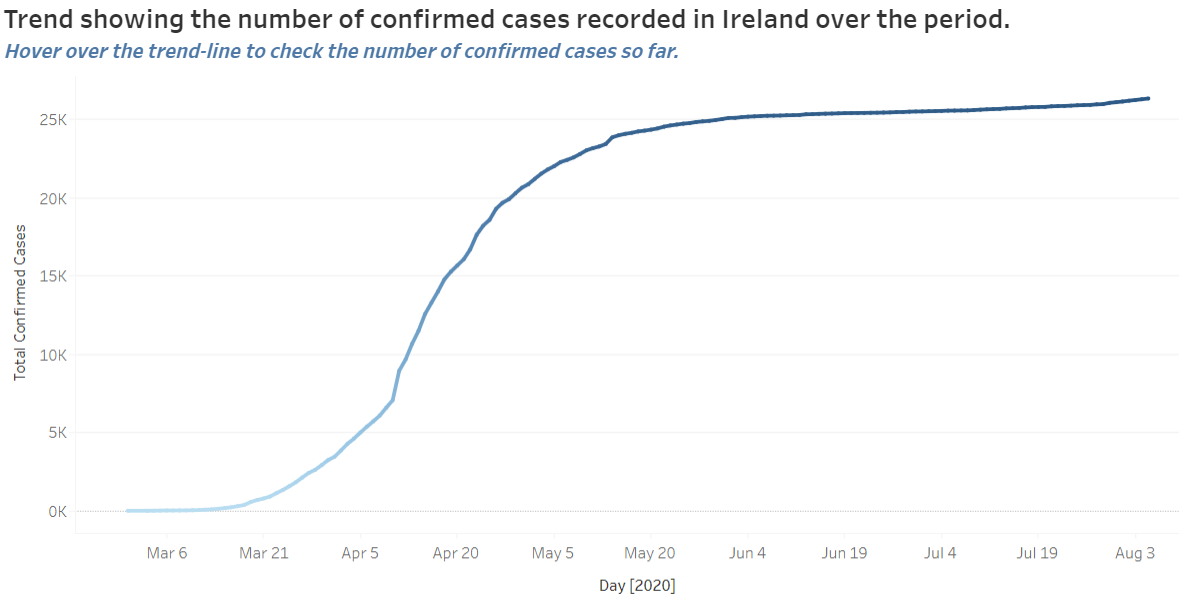


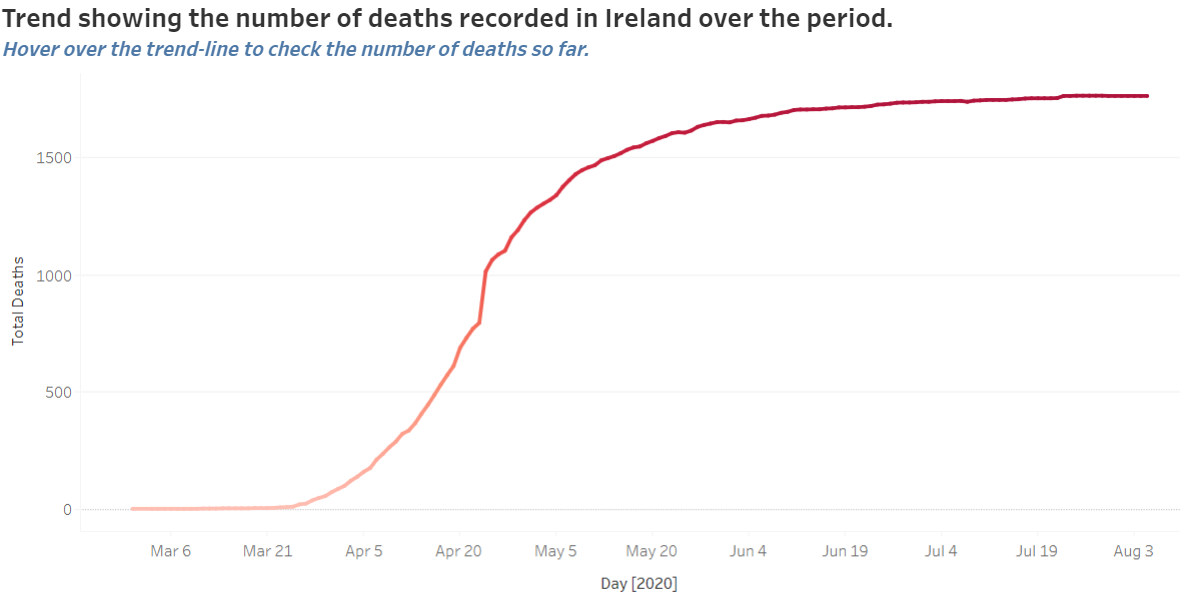












External Links!!

1. <https://api.covid19api.com/countries> (Country Details)

2. <https://api.covid19api.com/country/south-africa?from=2020-03-01T00:00:00Z&to=2020-04-01T00:00:00Z> (Country Status Details)

3. <https://api.covid19api.com/total/country/china> (Total stats/details per country)

4. <https://api.covid19api.com/all> (Everything about COVID-19)

5. <https://api.covid19api.com/summary> (Summary of stats like total cases, recovered cases, deaths)

6. <https://github.com/CSSEGISandData/COVID19/tree/master/csse_covid_19_data/csse_covid_19_time_series>  (Global Data (Tallied records: Total cases, Recovered, Deaths) Time Series Data)

7. <https://www.programmableweb.com/api/yatko-coronavirus-map-rest-api-v10>

<https://api.quarantine.country/api/v1/summary/latest> (Yatko API)