CSCE 5320

Scientific Data Visualization

Data visualization on COVID 19 Pandemic

Submitted by

Likith Guduru

Naveen Paritala

Sai Sandeep Gollamudi

Dinesh Tadepalli

Goals and Objectives

Motivation

During Covid 19 pandemic new challenges arose every other week, there has been first wave which we thought was the last one then came the second wave. There have been a lot of situations, problems because of different variants across the globe. Older people were the most impacted group of all.

There are different scenarios and reasons behind the ravage that the pandemic has caused, and each country has its own reasons behind the load of cases and deaths. So, we as a group found these factors to be compelling and would make a good project.

Significance

Pandemic is something that perhaps occurs once in a lifetime, or it won't be exaggeration to call it's a once in generations unfortunate event that no one has imagined would happen. There are a lot of health-related issues and conditions, but pandemic is something different study altogether. There are different parameters that can be analysed like age, GDP, median age of the country, vaccination rates, test load, virus reproduction rates. There are different comparisons that can be drawn among the nations depending on factors like density, area of the country, population etc. So, it is important and significant that we study and explore the data of different age groups and risk factor like death for that age group, work and analyse more on vaccination rates and then compare it with Covid test cases rate to find if any correlation exists. It is also important to compare the data of two countries with similar parameters.

Objectives

We want to compare the Covid cases and death tolls in two countries with a similar population that fared differently during the Covid pandemic. We as a team, using the covid dataset want to explore and highlight the underlying reason that caused the difference in the reported cases and deaths between the countries.

Features

The dataset for this project has been taken from Kaggle, there are many parameters, and we would like to choose the most important parameters that can make meaningful visualizations using Tableau, Python and HTML. Using those parameters, once different visualizations are drawn, we would embed few visualizations in web pages and host the website.

Introduction:

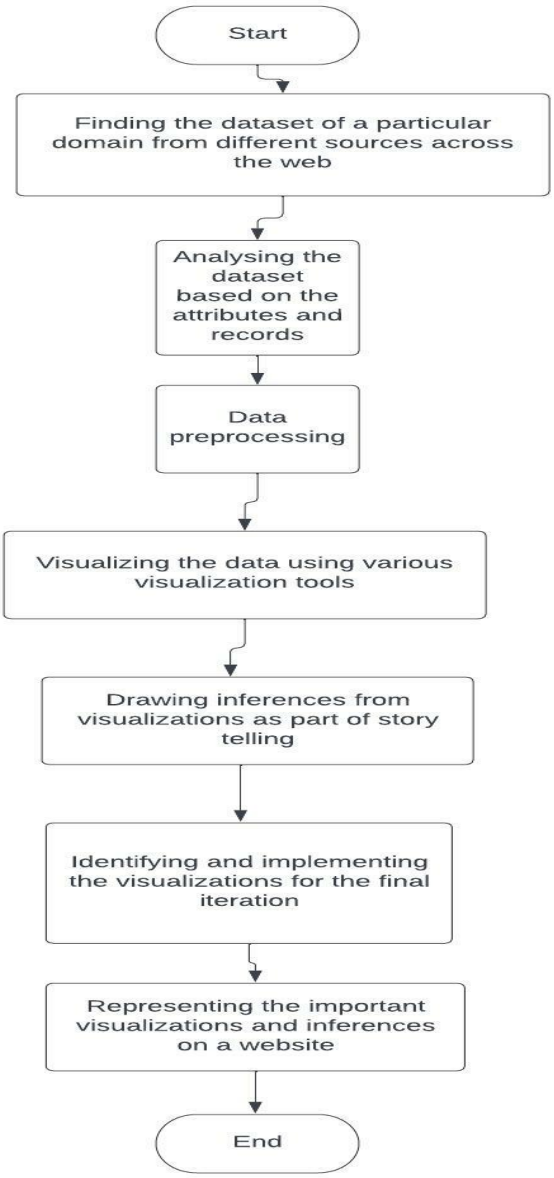
Domain

We are working on Covid 19 pandemic dataset. It comes under healthcare domain. However it involves various factors and isn't exclusively Healthcare domain. Various factors like socio, economic, geographical conditions have a direct or indirect effect on Covid 19 induced pandemic. So, It can be considered a special case in Healthcare domain as this is also a once in a century phenomenon. The Visualizations produced and inferences drawn can be used by various groups. Healthcare analysts, researchers and other communities of the domain and even out of domain stakeholders can make use of the visualizations that have been produced.

Workflow diagram with explanation:

The workflow diagram for our project is shown below. This workflow diagram gives visual analysis on how we progressed in our project through various phases.

We have searched for various datasets through the internet and finalized this dataset in health domain as this is the ongoing situation globally. We selected the dataset by browsing various sources through the web. After selecting the dataset we analysed the dataset for each attribute and record and made necessary pre-processing tasks required for the dataset. We then made the visualizations using various visualization tools like tableau, python and HTML for our story and drawn the conclusion.



Related background work:

Covid-19 is a virus that was first occurred in November 2019 in Wuhan, China and later it was widely spread globally. Most of the people came to know the word pandemic after the existence of covid. We have gone through various datasets through Kaggle and websites but unable to find any appropriate one until we selected this dataset.

https://www.tableau.com/covid-19-coronavirus-data-resources

In the above article, there are various visualizations on Covid 19 related data, we first came across the above page and upon looking at the work that has been done using different parameters for different scenarios like cases, deaths, positivity rate, vaccine rollout, vaccines administered, tests and various other scenarios, we understood that the scope for performing data analysis on Covid 19 induced pandemic is great and we then decided to go ahead with this domain. Using the visualizations provided in the above article we got an idea on how to move forward using our dataset. Though we didn't perform the visualizations used in the

above article, it helped us gain the confidence on moving ahead with this domain which is vast and keeps evolving with time due to various real world scenarios.

Data Abstraction:

- **Dataset (Type and Attributes):**

SL.NO	Attribute	Datatype
1	Continent	Categorical
2	Location	Categorical
3	Year	Ordinal
4	Total deaths	Quantitative
5	Total cases	Quantitative
6	Total deaths per million	Quantitative
7	Population	Quantitative
8	Population Density	Quantitative
9	Age 65 or older	Ordinal
10	Age 70 or older	Ordinal
11	Life expectancy	Quantitative
12	Male_smokers	Quantitative
13	Female_smokers	Quantitative
14	Extreme poverty	Quantitative
15	GDP_per_capita	Quantitative
16	Median_age	Quantitative
17	Reproduction rate of virus	Quantitative
18	Total vaccination	Quantitative
19	Cardiovascular_death_rate	Quantitative
20	Diabetes_prevalence	Quantitative
21	Hospital_beds_per_thousand	Quantitative
22	Human_development_index	Quantitative

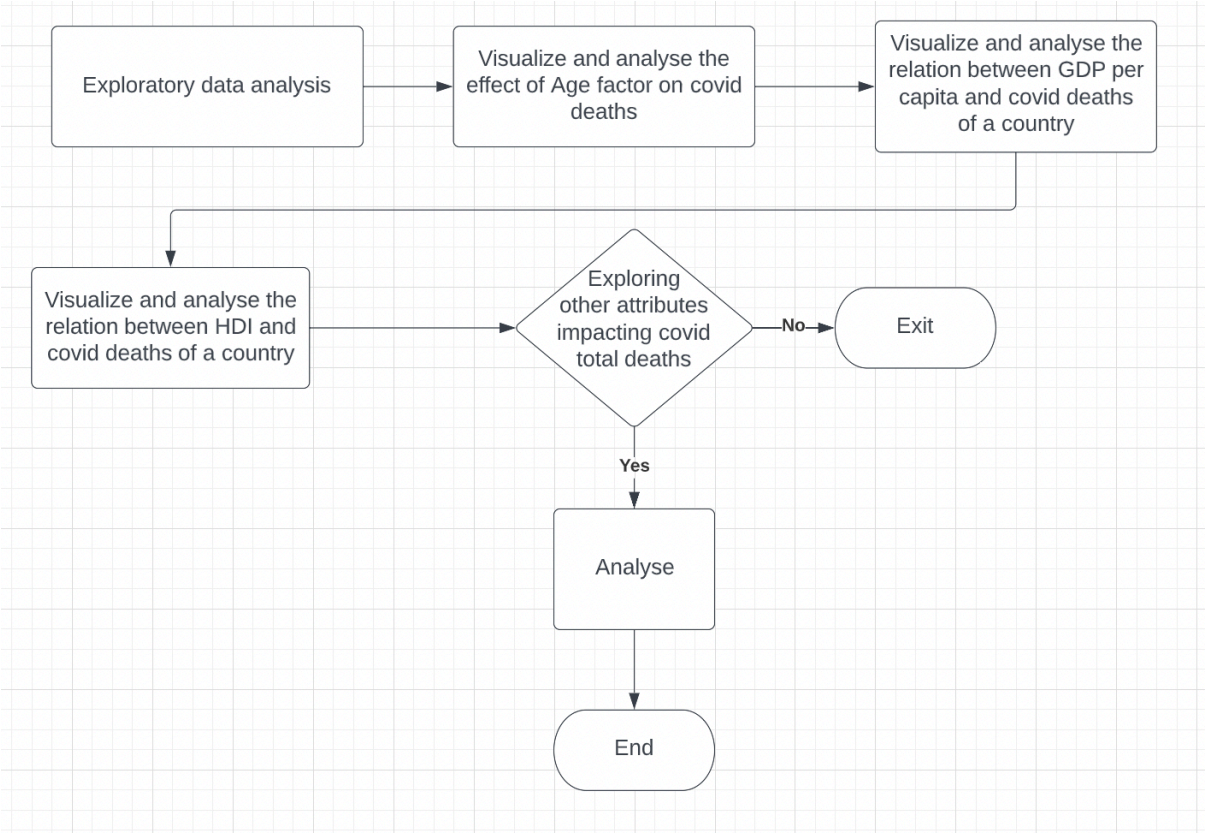
- **Detailed description of the dataset:**

  The dataset was made by the author in collaboration with the university of Oxford. Since the beginning of the covid-19 several researchers have collected data from number of sources and have developed a repository of various datasets on covid-19. This dataset however is updated every week on a regular basis since the beginning of the pandemic which describes various attributes that are linked to the covid-19 pandemic. Some of the important attributes that have been used for the part of this project were mentioned above in a tabular format. The description of the above attributes can be found below.

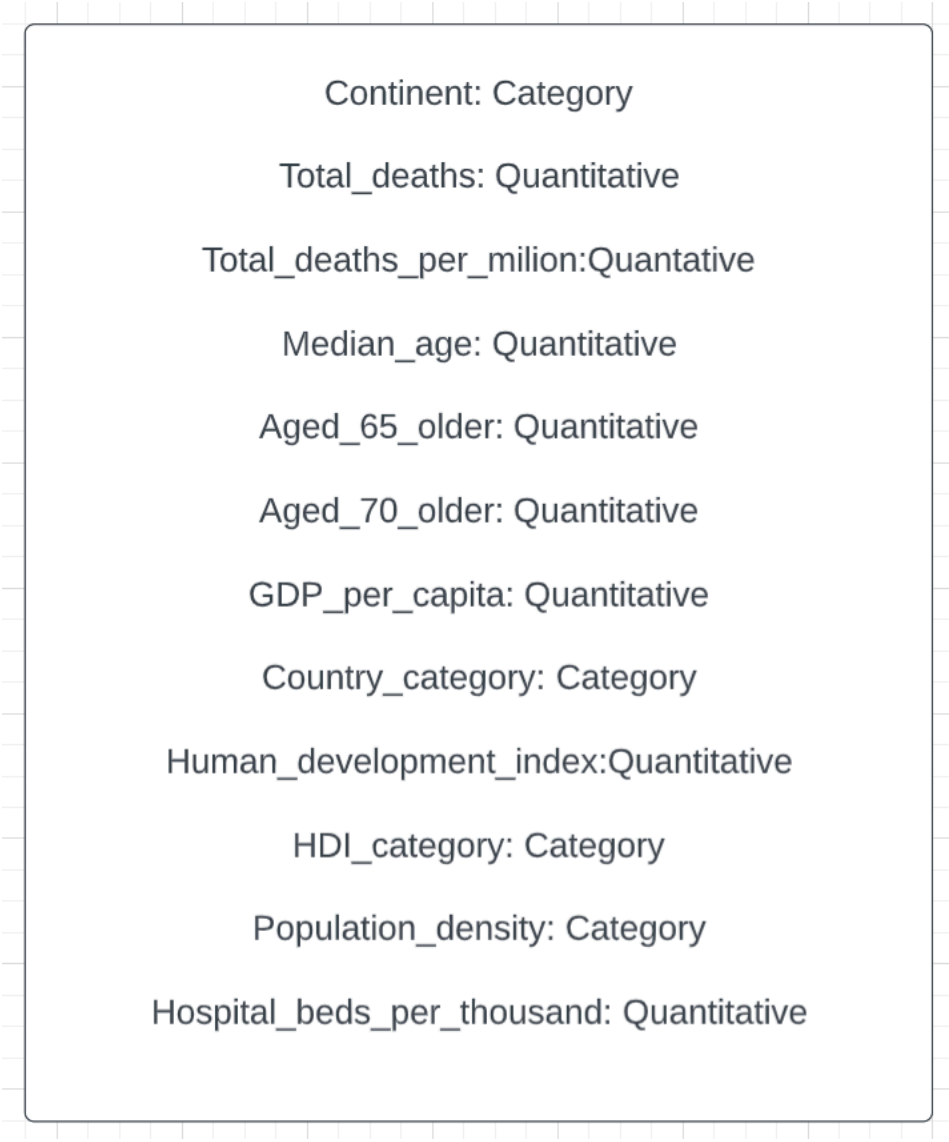
  Continent: Describes in which continent the country belongs to.
  Location: Location describes the country in which the covid cases have occurred.
  Year: Describes the year in which the cases have started and ongoing.

Total deaths: Tells the number of deaths occurred.
Total cases: Tells the number of cases occurred in every country.
Total deaths per million: Count of the number of deaths per million.
Population: Gives the count for the number of people in every country.
Population Density: Gives the count for the number of people living within a particular area.
Age 65 or older: The number of people with age 65 or above.
Age 70 or older: The number of people with age 70 or above.
Life expectancy: The average expected age limit that the people can live.
Male smokers: The number of males who are smoking
Female smokers: The number of females who are smoking.
Extreme poverty: The percentage of people of the nation living under extreme poverty
GDP per capita: The GDP of the nation with respect to the population. It is a crucial metric in determining the economic of the country.
Median age: Describes the median age of the country also describes how young the country is.
Reproduction rate of virus: Describes the spread of the virus to each individual.
Total vaccination: Gives the number of vaccines taken by the people in each nation.
Cardiovascular death rate: The count of people that die due to heart diseases per 1 lakh.
Hospital beds per thousand: The number of beds available in hospitals per 1000 population.
Human development index: Human development index which varies from 0 to 1 describes the overall well being on various parameters like life expectancy, education, health care facilities and other factors which include economic criteria.

- **Design of features diagram:**



The above picture describes the structure of our project and the design of features we are using is this project is mentioned in the below picture.



- **Data transformation:**

  We have transformed the dataset into a meaningful one by cleaning it, removing the null values and other unwanted records which makes it a proper useful dataset. The attributes that do not have any importance and significance have been removed. Initially there are records upwards of 1,66,000 since the dataset has countries more than 230, and has been updated once a week since the beginning of the pandemic way back in 2019. We took a cut off date of Feb17th, 2022 which was the last available date when we decided to proceed with this dataset for the project. Since there are multiple records for each country, we went ahead with the above-mentioned date for getting the cumulative values in a few attributes. We then performed data pre-processing and then have taken only the attributes that are required for visualization analysis and performed the necessary tasks for the project.

Task Abstraction:

- **Task(Target and Actions):**

  The initial focus of this project is to show the different behaviour of two countries with similar population and to recognize the underlying reasons. We have selected South Africa and Italy as the two countries to make the comparison and analysis as they are having similar count of population. This project insights on how the countries coped during the covid crisis and possible reasons for the difference in performance in between the countries. Furthermore, as the we progress we analysed the behaviour of all the countries across the world by taking the different parameters to compare and visualization. The analysis for this project is done based on the country's population, age factor, death percentage, GDP of the country, population density and so on.

  For the targets below visualizations have been implemented in the below sections of the document.

  Target 1: Analysing Data distribution of Covid "Total death" attribute for countries.
  Action: Used rug, KDE and Box plot in analyzing data distribution of Covid deaths parameters.

  Target 2: Analyzing the impact of age factor on Covid death.
  Action: Using joint plot visualized the relation between countries median age, aged 65 or older, and aged 70 or older attributes with covid "Total deaths per million".

  Target 3: Effect of GDP per Capita on Covid deaths.
  Action: Using strip, violin, box plot visualized the relation between "GDP per Capita" and Covid's "Total deaths per million" attribute.

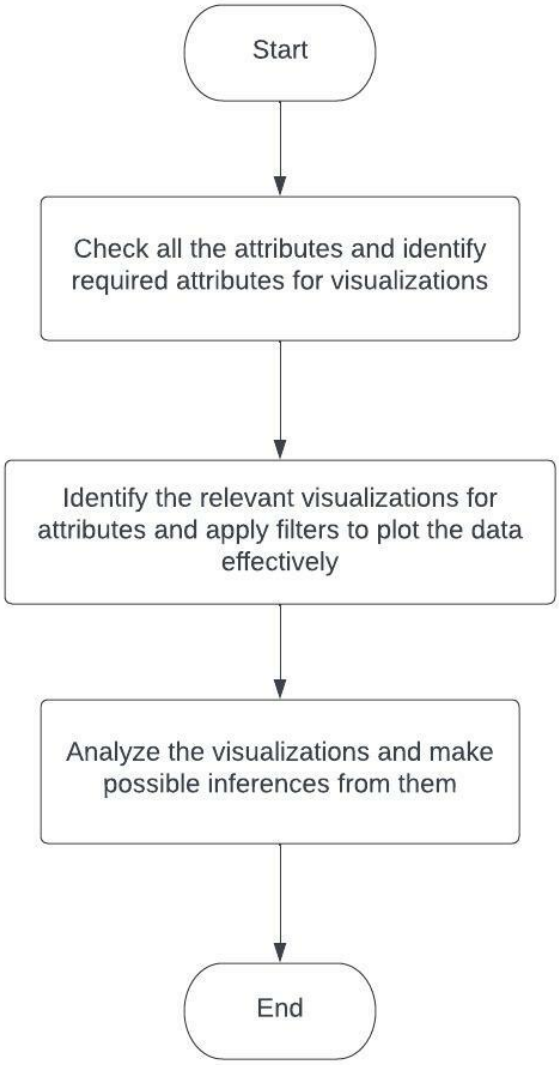
  Target 4: To visualize the possible effect of Human development index on covid deaths.
  Action: Using join, strip and violin plot visualized and analysed the effect of HDI on the "Total death per million" attribute.

  Target 5: Explore other attributes' effects on covid deaths.
  Action: Using bar, lm plots explored other attributes' effect on country's covid "deaths per million" attribute.

  Workflow diagrams with explanation:

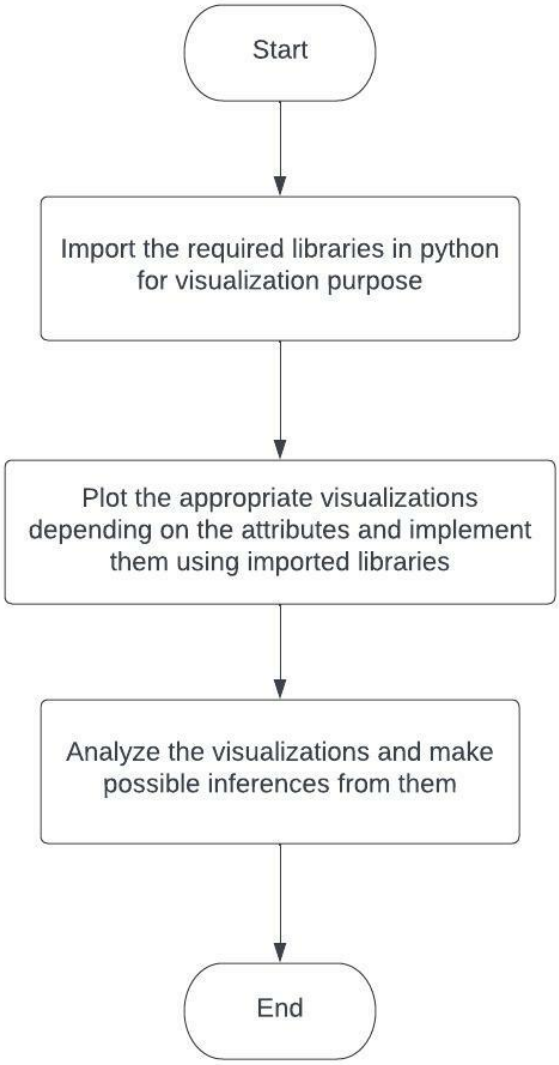
  We implemented visualizations using tableau, python and HTML and CSS. For implementing visualizations using tableau the procedure is to first identify the attributes from the dataset and them choose an appropriate visualization technique and then plot the graph using appropriate filters. Once the visualization is found to be effective and relevant, analyse the plot to make any possible conclusion.

Tableau Workflow



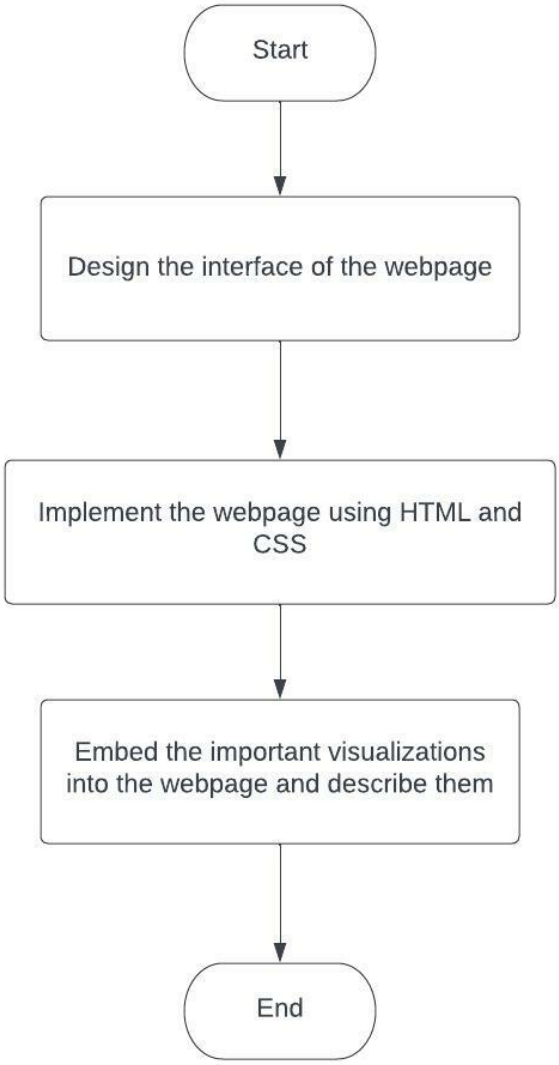
For implementation using python, we first have to analyse various visualization libraries in python and then choose the libraries which we think would be useful for plotting the graphs according to the attributes in our dataset. Once the libraries are imported we then have to plot the graphs using those imported libraries and analyse the plots.

Python Workflow



For implementation using HTML and CSS, we first need to finalise the webpage user interface and then implement the webpage upon which we need to embed the important visualizations, descriptions and possible conclusions if any.

Webpage workflow



Implementation of tools used:

For visualizations in this project, we have used Python, Tableau and web technologies like HTML, CSS.

1) Python has a set of libraries specifically for visualization purposes like Seaborn, Matplotlib, Altair etc. For the implementation of this project we used Seaborn, Matlplotlib libraries. **MatplotLib** is a comprehensive library for creating static, interactive visualizations in Python. It is a plotting library. Seaborn is a data visualization library that is based on Maltplotlib. It is effective in using and implementing arrays and dataframes. One can make attractive graphical

representations using seaborn library. We used both these libraries for visualization purposes in Python.

Pros:

- Widely used in data science community.
- Python supports multiple libraries for visualizations and is open source.
- Easier in debugging as it executes the code line by line.

Cons:

- Performs a little slower when visualizing huge number of records.

2) Tableau

Tableau is a software used to view, understand, analyse data. It allows the user to connect to databases and make meaningful visualizations of the data and inferences can be drawn upon analysing the data. It is also used for business intelligence purposes. It is quick and easy to use. We used tableau software as part of the project for a few visualizations that are made from the Covid 19 dataset.

Pros:

- No prerequisite knowledge on computer science and coding is required.
- It is very user friendly.
- Despite of being easy to use and requires no prerequisite knowledge it is high performance tool and provides results in fraction seconds.
- One of the widely used and accepted tool in industry.

Cons:

- Manual effort is needed from the user end.
- Automatic refreshing using scheduling is not possible in Tableau.

3) **HTML & CSS**

We developed a webpage using HTML & CSS and put together the important visualizations that we have performed, in the webpage. The webpage also describes about the overview of our project, mentions about some key visualizations and explains the significance and inferences drawn upon them.

Pros:

- Supported by every browser.
- It is open source and can handle any visualization.

Cons:

- Doesn’t have any pre built charts.

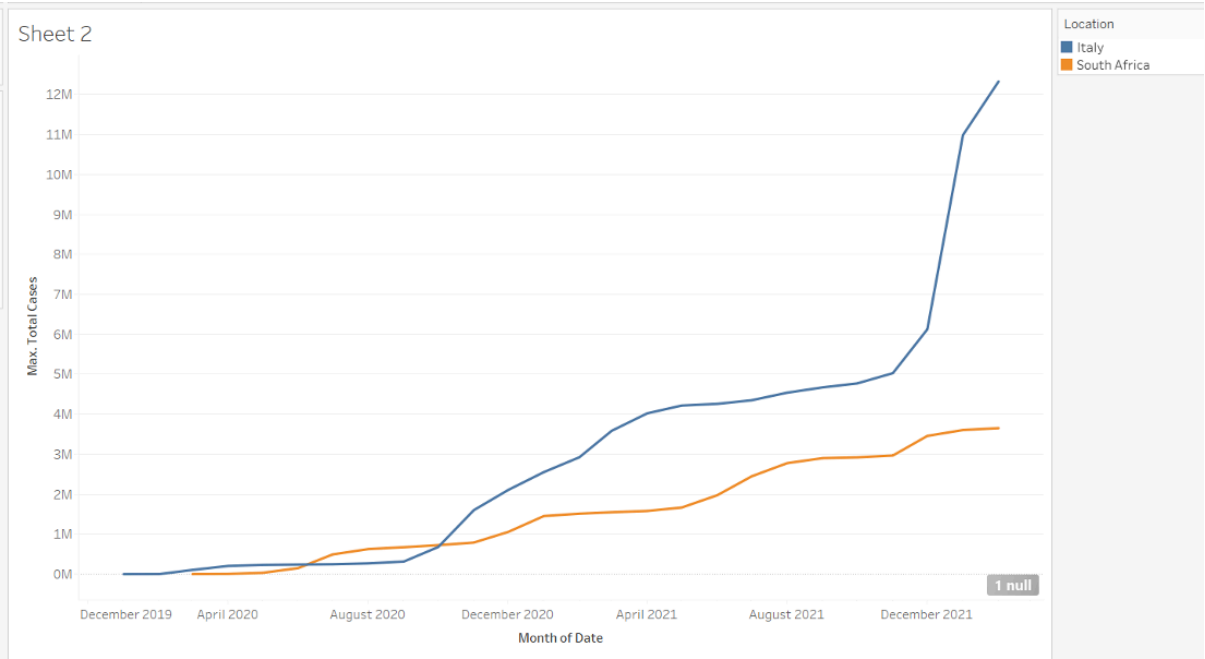
Results for Analysis:

- **Visualization graphs with detailed explanation:**

  The initial idea of this project is to compare two countries with an almost similar populations, but which fared differently during the COVID pandemic. We analysed various countries to find the ideal pair with a similar population and recorded different results during COVID. We came across Italy and South Africa, both countries have a population of 6,03,67,471 and 6,00,41,996 but there is a significant difference between the total covid cases and deaths reported in these countries.

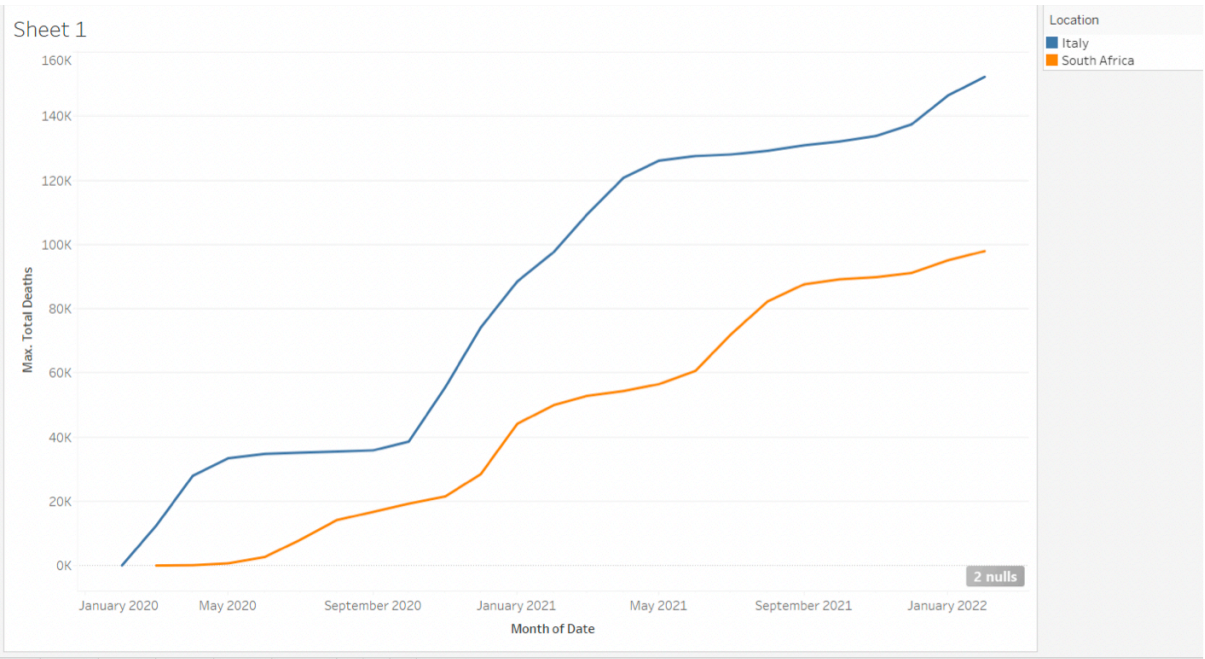
  The histogram of both the country population is shown below.

As mentioned earlier the country's population are close, but the reported cases vary significantly. Below are the reported cases and deaths in Italy and South Africa.



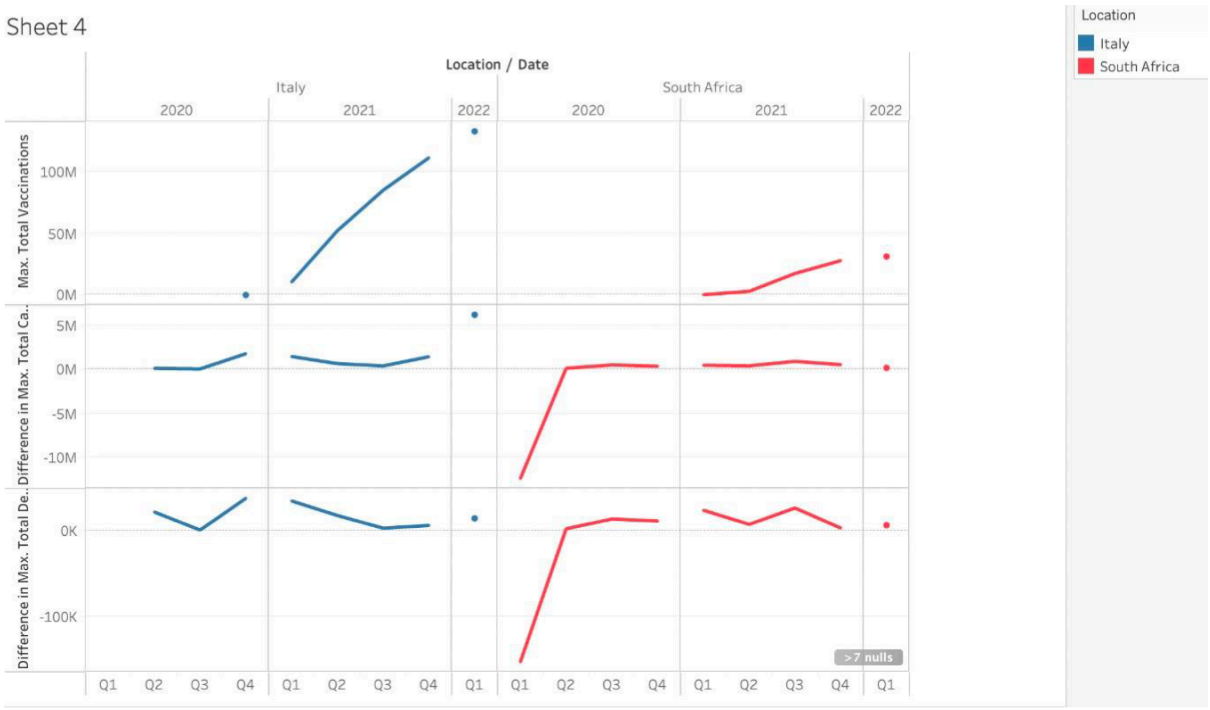
Total covid cases reported in Italy and South Africa from Jan 2020 to Feb 2022

The Covid cases started getting reported from Jan 2020 in Italy, whereas in South Africa the cases started getting reported in March. The cases started increasing in both the countries from September but the reported cases in significantly higher in Italy than in South Africa, which is clear from the above line graph. The pattern continued until the end, Italy has 1,23,23,398 by Feb 2022, on the other hand, South Africa has total cases of 36,52,024.



Total deaths reported in Italy and South Africa from Jan 2020 to Feb 2022

In Italy, the total deaths are reported in Feb 2020, whereas in South Africa the deaths started happening in March. Even though the deaths in both countries increased, In the end, the total death in Italy is 1,52,282 whereas for South Africa it is 97,955.

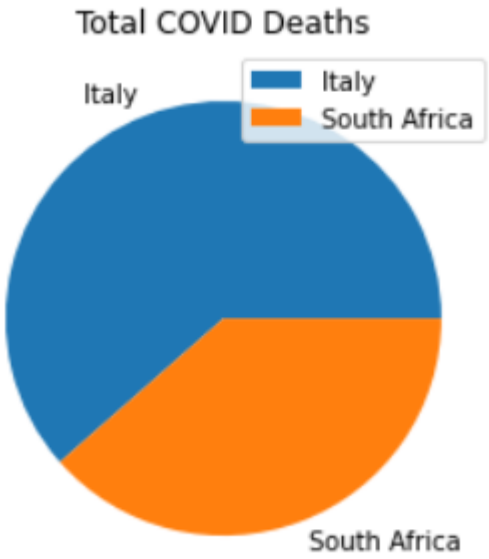


Corelation between vaccination, covid cases and covid deaths

Before the vaccination there was a huge increase in number of cases and deaths. As the vaccination program started in the start of 2021 in both the countries, we observed that the cases graph stabilized in both the countries. Whereas the covid death curve plummeted in Italy whereas it fluctuates for South Africa.

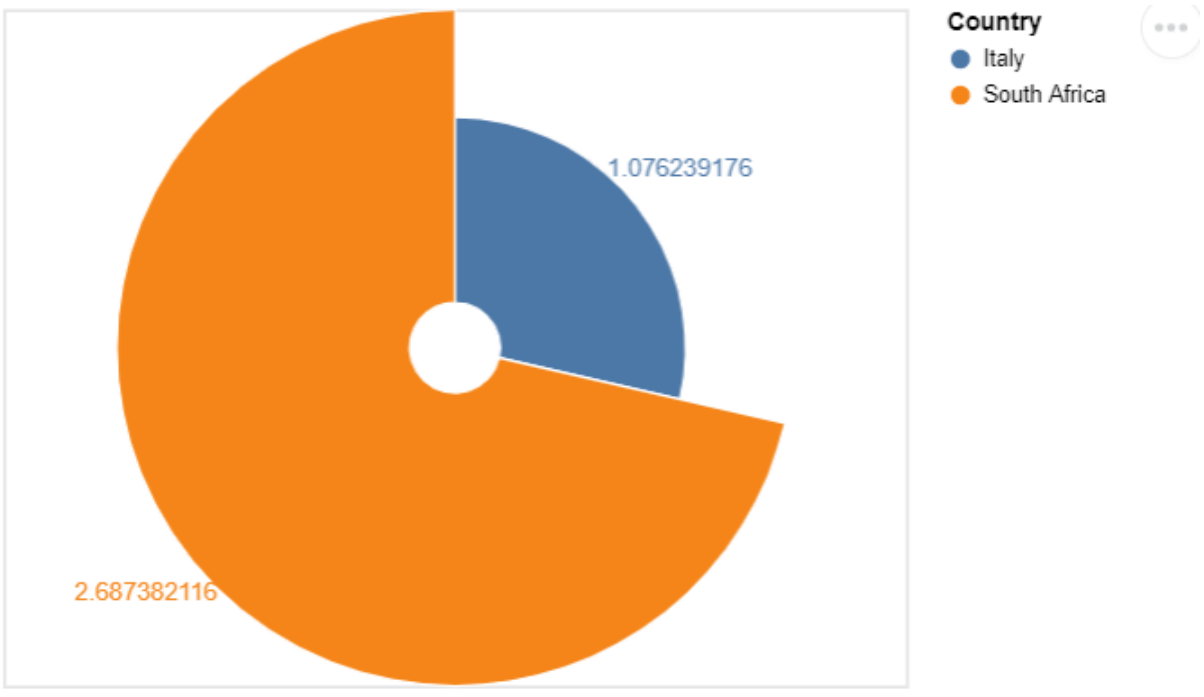
Coming to the reason, why we've used histograms and line graphs for the visualizations most of our attributes are of Quantitative data type. So, histogram serves the purpose better than rest of the visualizations. We also have a few data which is collected over the time, for these visualizations we've used line graph as it is best in conveying the data comparison over time.

In increment-1 we worked on different parameters that might have affected the covid deaths and cases in Italy and South Africa. We have used Tableau to construct visualizations in increment 1. So, for increment-2 we decided to construct visualizations using python libraries like Matplot, seaborn and Altair. We are continuing from where we stopped in increment 1, below is the pie chart for total covid deaths between Italy and South Africa. We can clearly observe that the number of deaths in Italy is significantly higher than the number of deaths in South Africa.



Pie chart for total covid deaths between Italy and South Africa using Matplot library

Now we are plotting the death percentage of covid between Italy and South Africa using altair library in python.

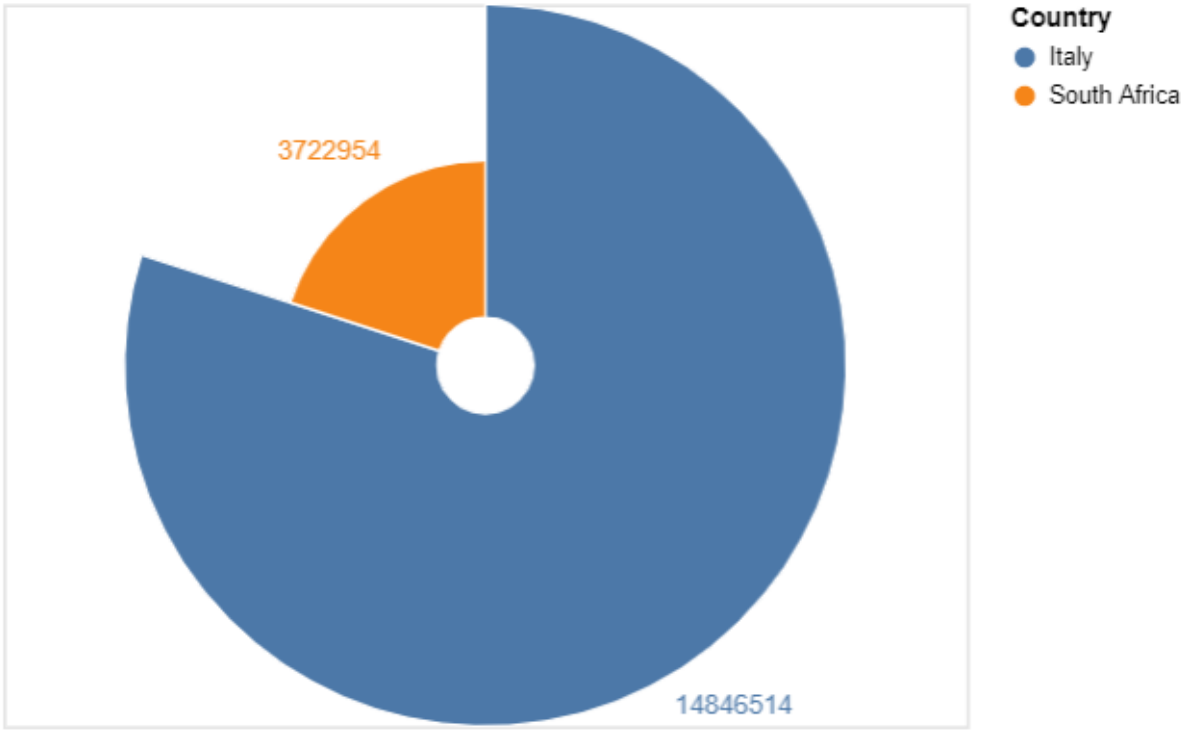


Pie chart for death percentage between Italy and South Africa using Altair library

The death percentage in Italy is 1.07 whereas the death percentage in South Africa is 2.68 as observed in the above radial plot. Here we can notice that even though the number of deaths

in Italy is higher than South Africa, the death percentage in Italy is less compared to South Africa.

We plotted another radial plot between Italy and South Africa, this time we plotted it for total covid cases using Altair library.



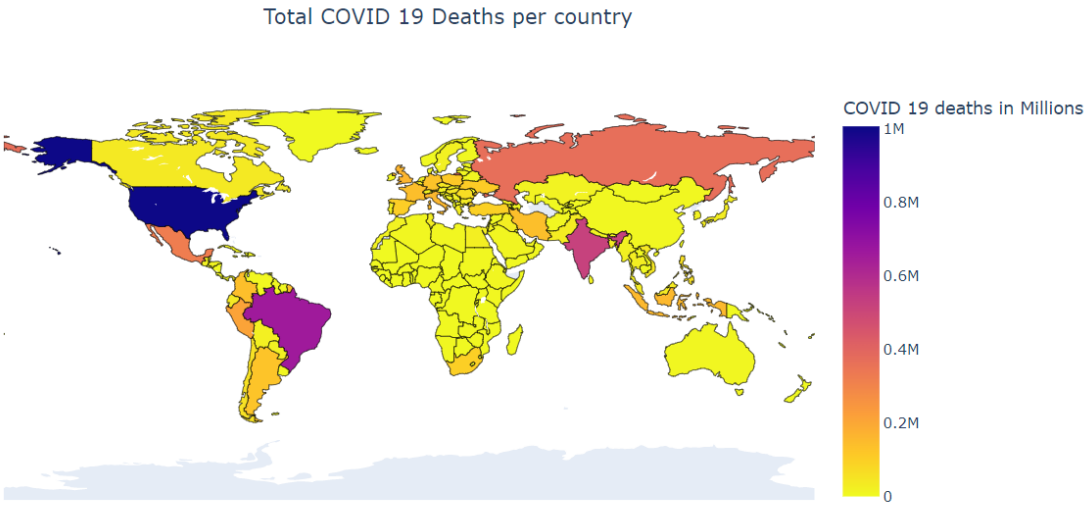
Radial plot for Covid cases between Italy and South Africa using Altair

As we can see in the above graph, the number of cases in Italy is near to one crore forty-eight lakhs whereas number of cases in South Africa is near to thirty-seven lakhs. Italy is highly affected by covid compared to South Africa, the number of cases in Italy is nearly 4 times higher than South Africa. As number of cases in Italy is higher, this is the reason why number of deaths is also higher compared to South Africa but when we compare the death percentage due to covid, the death percentage is higher for South Africa than Italy.

We explored all the possible factors that might have affected the possible covid cases and death difference between Italy and South Africa. Now we want to analyse and see if the same reasons are affecting the covid deaths globally.

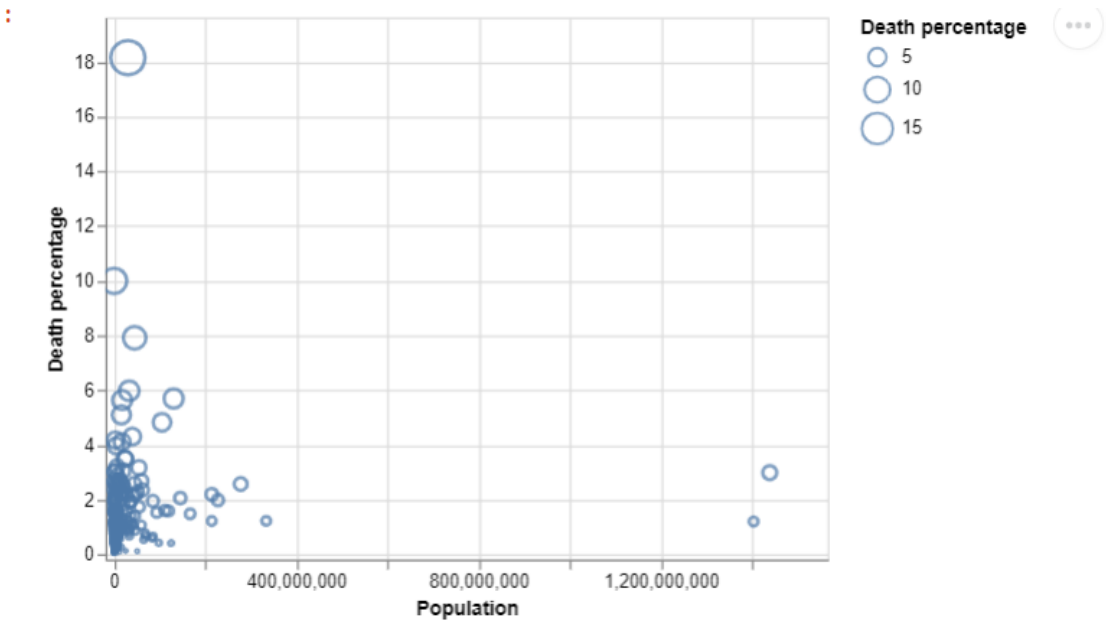
Global Covid death analysis

First, we want to plot total covid deaths in each country on the world map to understand if there any areas which recorded Covid deaths in the large number. If we can identify such areas, then we can explore what factors are common in those regions and analyse them. The total Covid deaths per country is plotted on the world map with the help of Plotly library in Python.

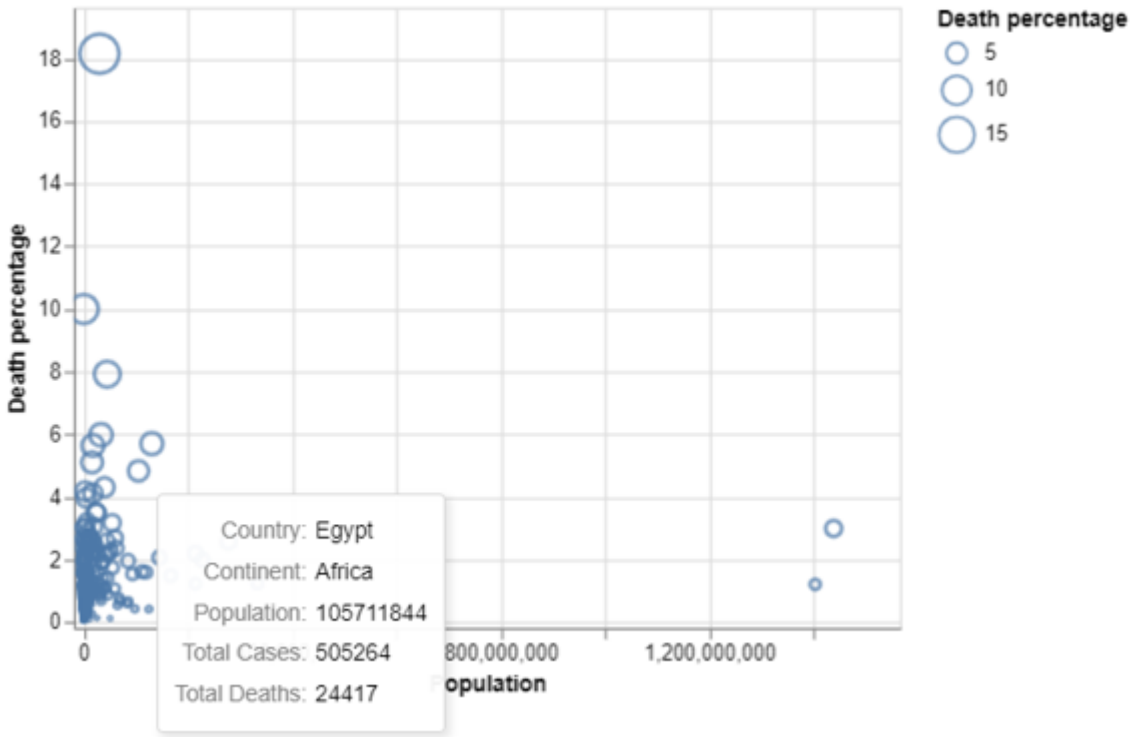


Total COVID 19 deaths per country plotted on World map using Plotly library in Python

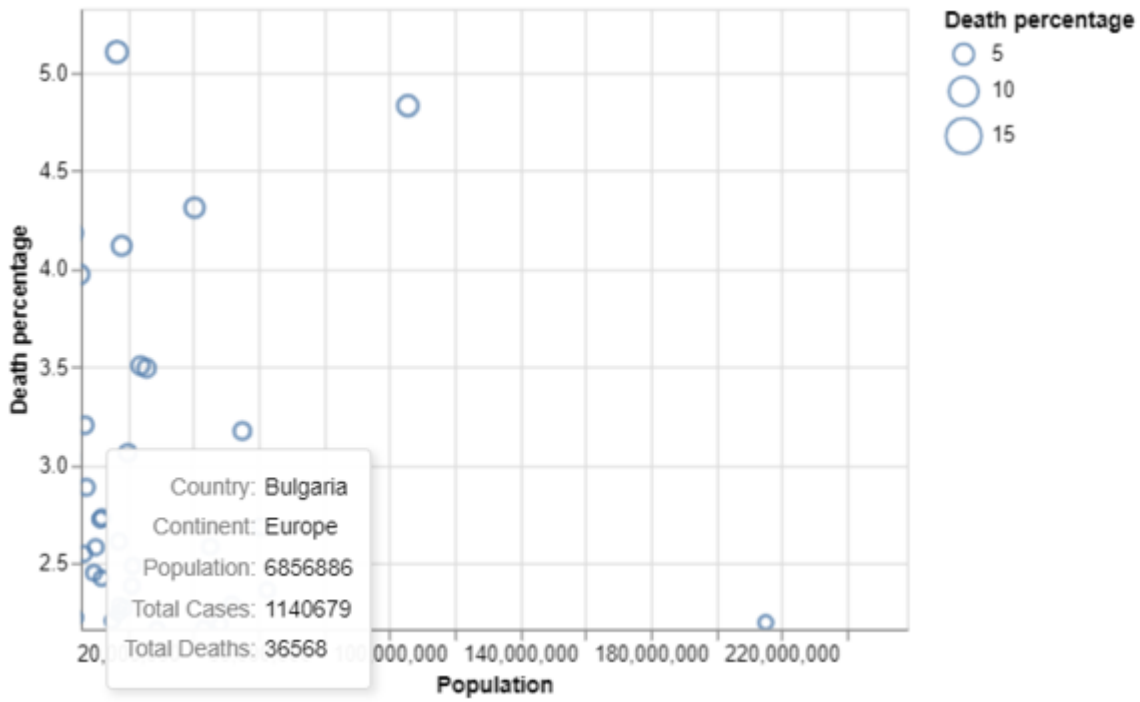
As we didn't identify group of countries in a zone that recorded severe covid death. We decided to plot a interactive scatter plot between Total deaths per million and Population. We can zoom and pan the scatter plot. By using this interactive functionality, we can further explore the specific area on the plot we want to observe. Furthermore, we added tooltip functionality in this plot, so we can hover on any particular point on the plot and we can view the name of the country, continent it is present in, population of the country, total cases recorded in the country, and Total deaths recorded in the country by covid. This functionality helps viewer gain even more required information about the country, Covid cases and deaths. This interactive scatter plot is plotted using Altair library in python.



Interactive Scatter plot constructed between Covid death percentage and population of the country using Altair.



Tool tip of Egypt point on the interactive scatter plot constructed using Altair

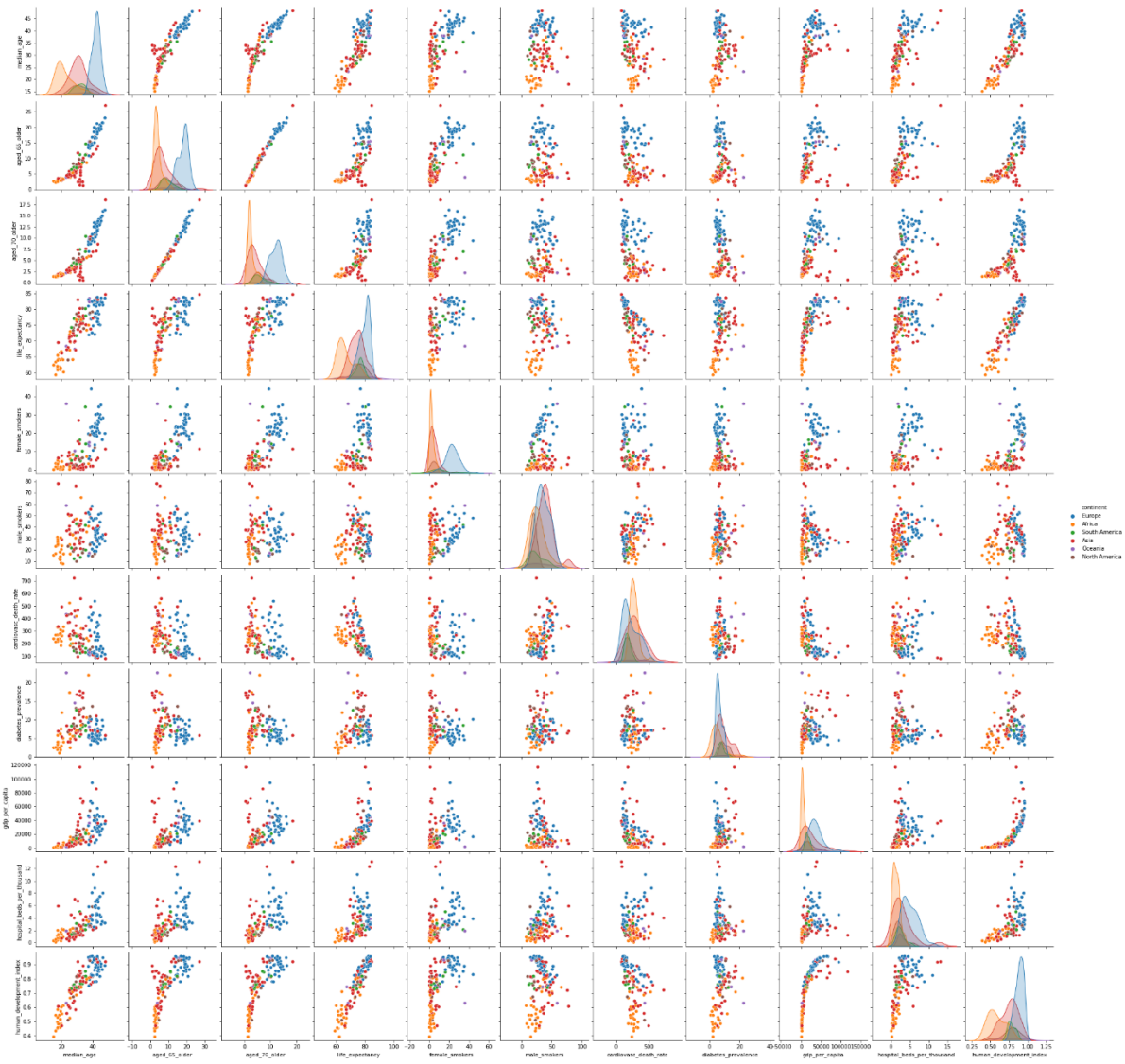


Tool tip of Bulgaria point on the interactive scatter plot constructed using Altair

We wanted to further explore different factors that might have influenced covid deaths all around the globe, so we have searched for dataset which have multiple indexes, parameters and values that are used in measuring a country standard. We found a dataset in Kaggle which has the daily New Covid Cases, Deaths, Total Cases and Deaths on that day along with the above-mentioned parameters like GDP per Capita, Human Development Index, Diabetes prevalence, extreme poverty, life expectancy and various other such parameters. This will a perfect fit for exploring different inferences and visualizing it.

The dataset consists of nearly one lakh sixty thousand record with sixty-seven attributes, as we only wanted total death and cases per each country along with the country's parameters and has nothing to do with daily case and death values. We have selected the last available record from each country in the exploratory data analysis and at the same time we selected few parameters from the available list and created a new data frame. By selecting required records and attributes, along with removing null values in exploratory data analysis we reduced the data frame size significantly.

Once the exploratory data analysis is completed, we plotted a pair plot between the quantitative data present in the data frame. We plotted pair plot as it gives us idea on relationship between two attributes, along with the distribution of single attributes. We can quickly get a idea on the trends and patterns in the data with just a glance. The pair plot is plotted using seaborn library in python.

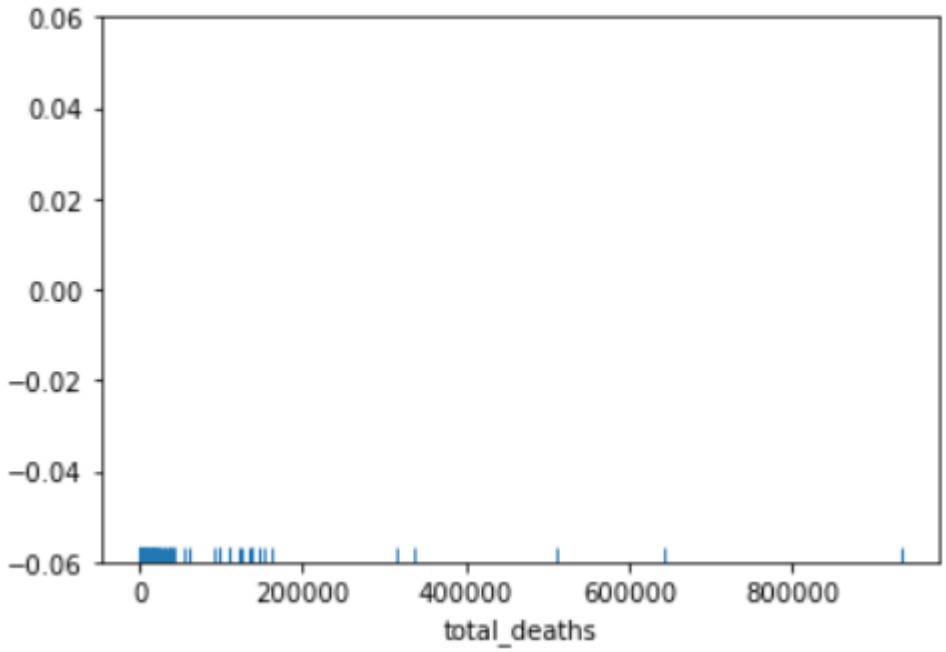


pair plot created using seaborn

The above created pair plot helped us in understanding the trends present in the data frame and the relationship between variables. Now moving into the analysis, we wanted to check the data distribution of total covid deaths in each country, to see the distribution of “total covid death” values. We decided to use Rug plot.

The definition of Rug plot is, it is used in plotting a single quantitative attribute to find its distribution. The value of that quantitative attribute is represented as blue marks on x axis.

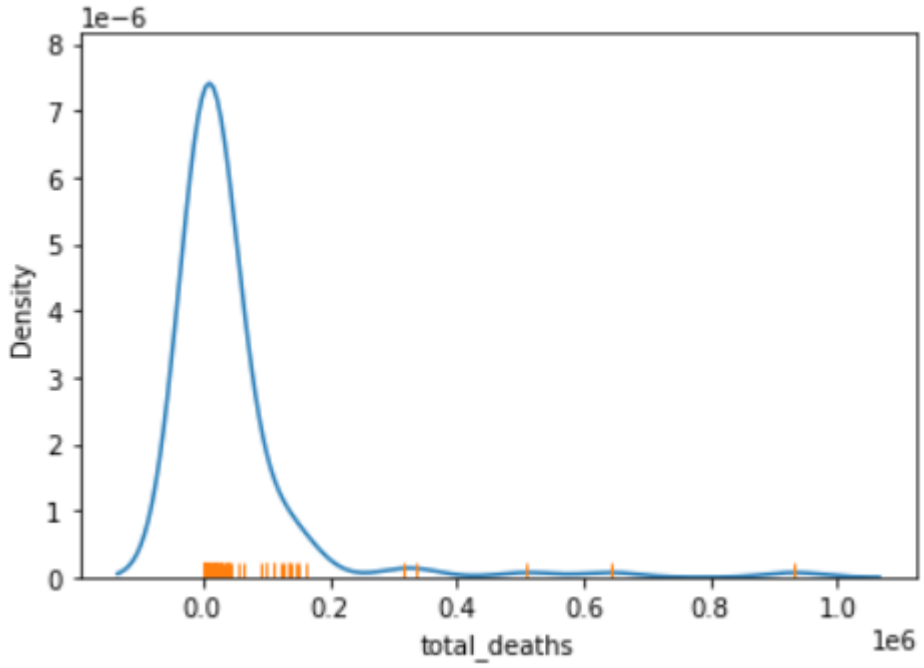
So, we created a rug plot using seaborn library in python.



Rug plot created using Seaborn for “Total deaths” attribute.

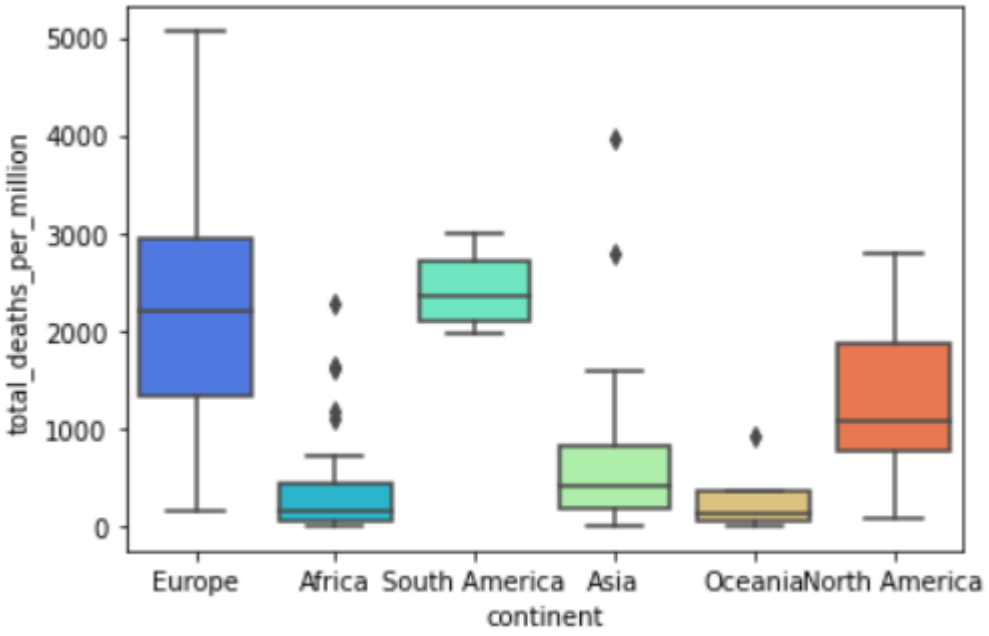
By looking at the above rug plot, we can notice that most of the death counts are high at one location on the x axis and moving to the right of x axis we can see the marks kept on decreasing. Once we crossed two lakhs mark on the x axis there is only five marks present on x axis, so there are only five countries which recorded more than two lakhs’ deaths. By looking at the rug plot we can come to the conclusion that for most of the countries have covid total death value less than a lakh and almost every country has less than two lakhs.

To make it further quickly understandable to the viewer we are combining the rug plot with the KDE(Kernel Density Estimate) plot. KDE plot plots the probability density function of the attribute selected, it is used in understanding the distribution of the data for that attribute. When we combine rug plot and KDE plot, the resulting plot will be much more quickly and easily understood by the viewer.



Rug and KDE plot created using Seaborn for “Total deaths” value.

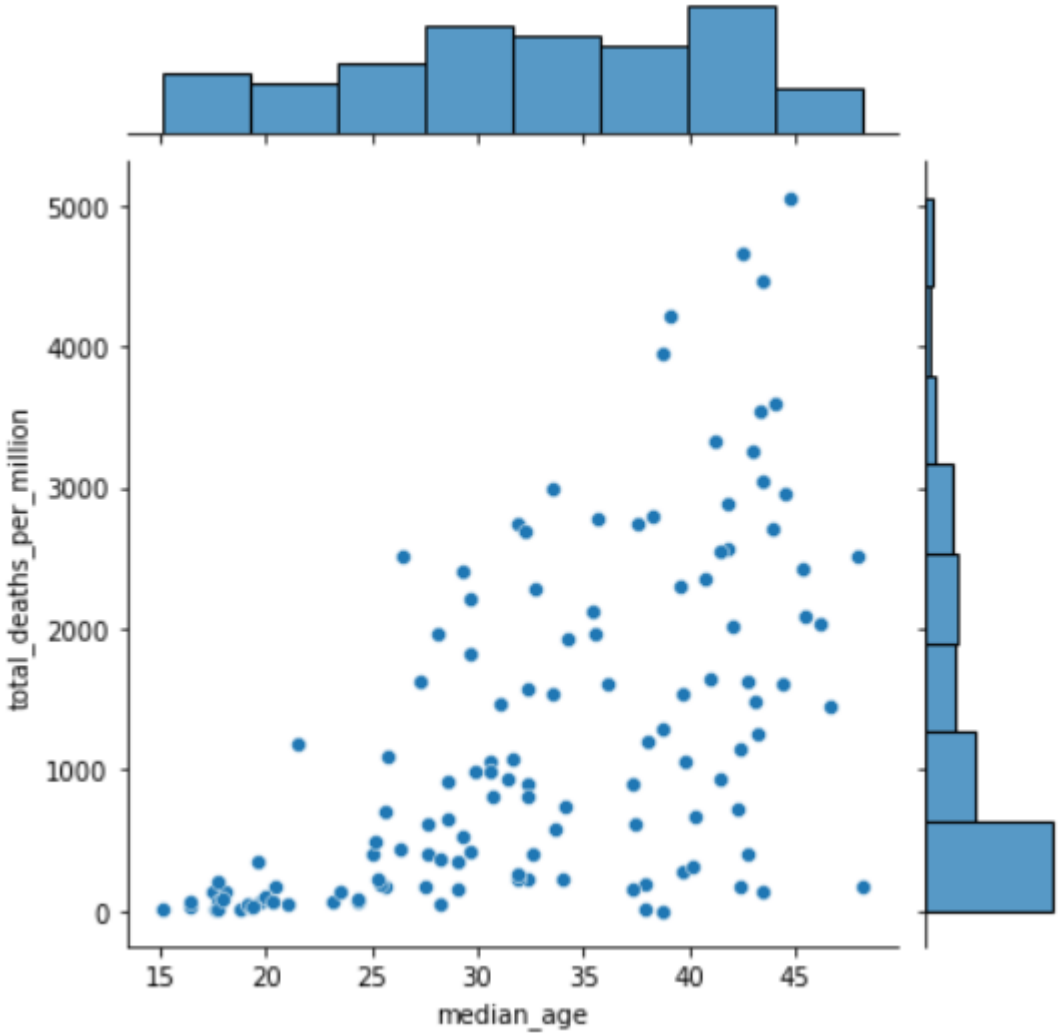
Now to understand how the covid deaths are distributed across continents we plotted a box plot, by using total deaths per million values on y axis and continent value on x axis. We used box plot as it gives information about outliers, median, interquartile range for each continent.



Box plot plotted using Seaborn

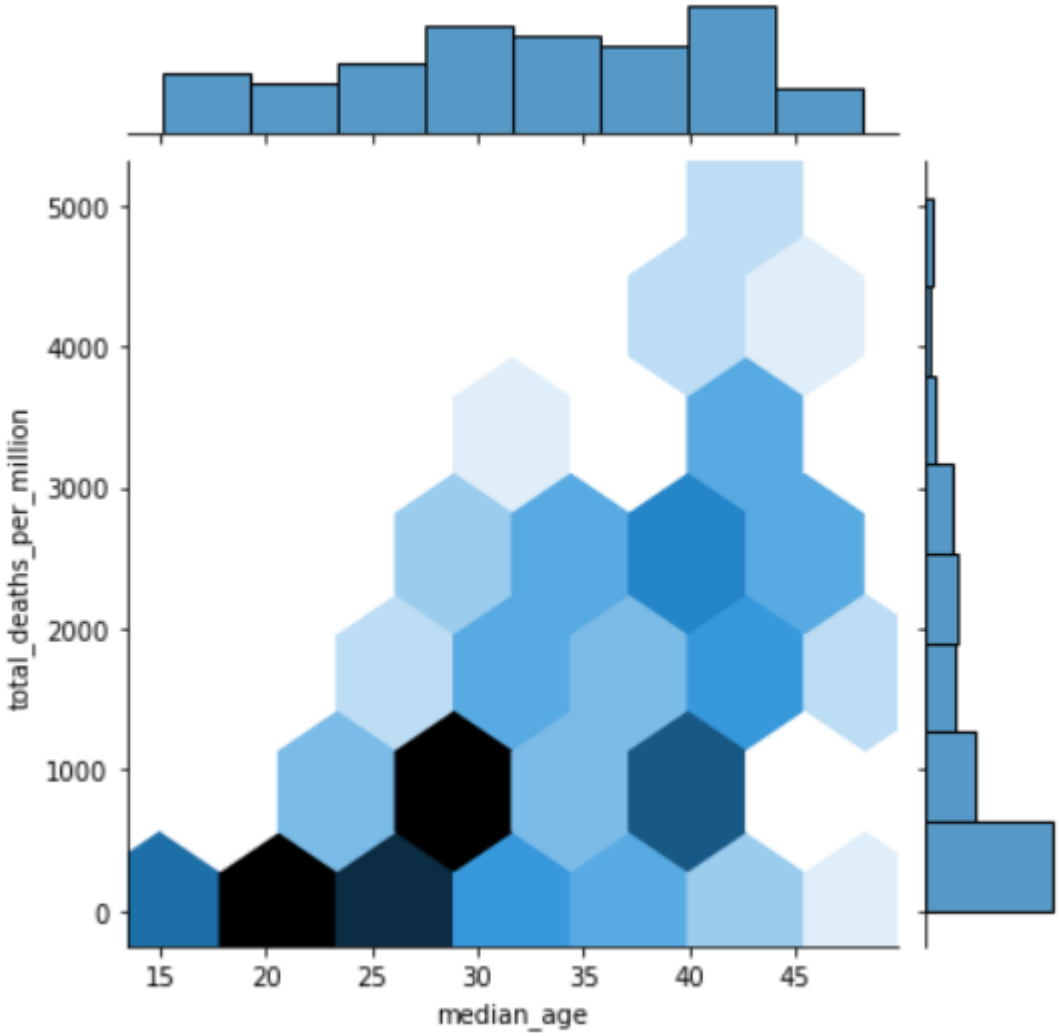
By looking at the above pictured box plot, we can observe that Europe has a largest interquartile range and max value. Both Europe and South America has a similar Median value. The Median is lowest in Oceania and Africa. This is the distribution of total deaths per million in each continent.

From the start of the covid pandemic, every government and research organization stated that covid is much deadlier for aged people. As different countries have different percentage of older population and different median age, we decided to see if there is any difference in between countries with low older population and countries with higher older population.



Joint plot constructed between the total deaths per million and median age using Seaborn

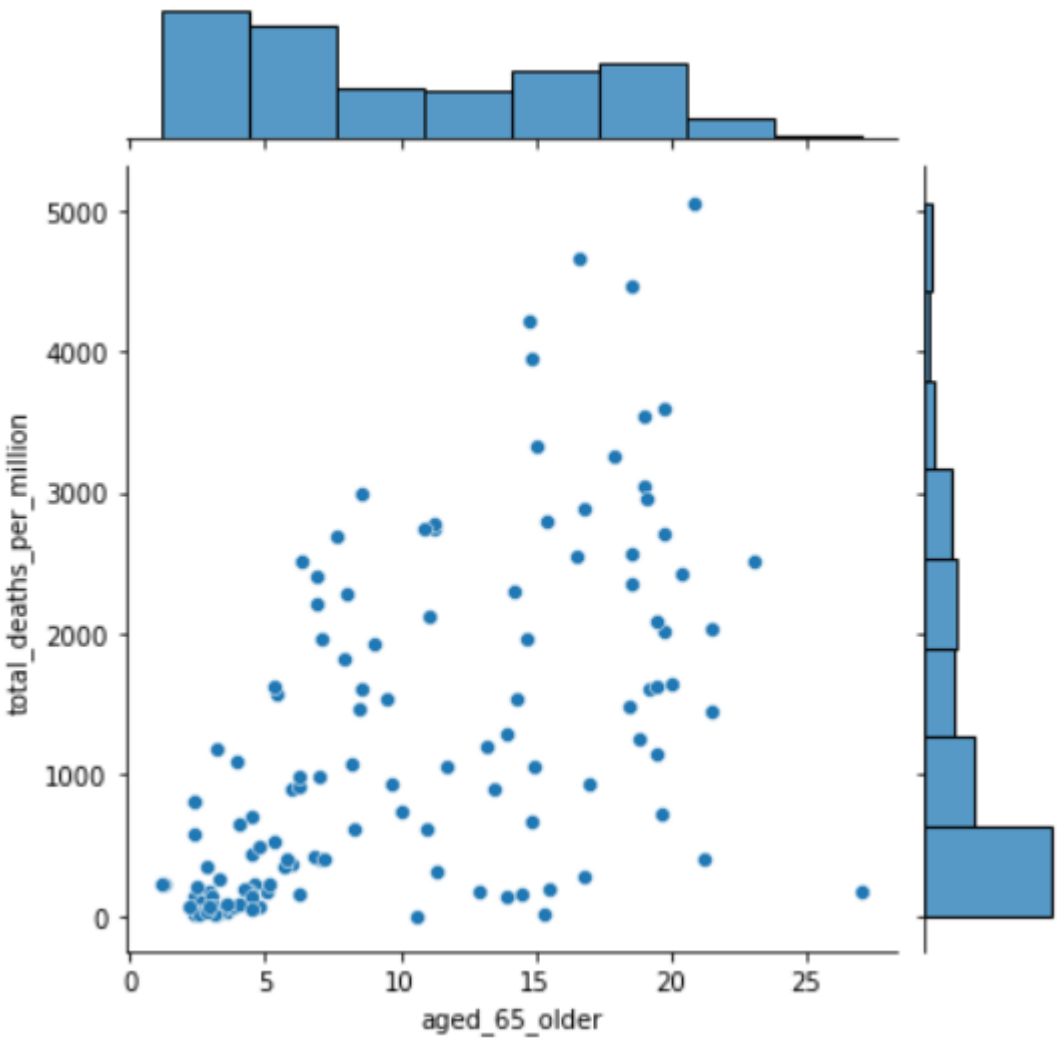
We can observe that countries with less median population has less total deaths per million but as we move towards the right side on x axis, as median age increase we can see the total deaths per million value gradually started increasing for some of the points. As we go to the end of the x axis we see total deaths per million reaching four thousand and five thousand values, though there are some records which doesn't have large number of deaths per million even though it has higher such records are far less.



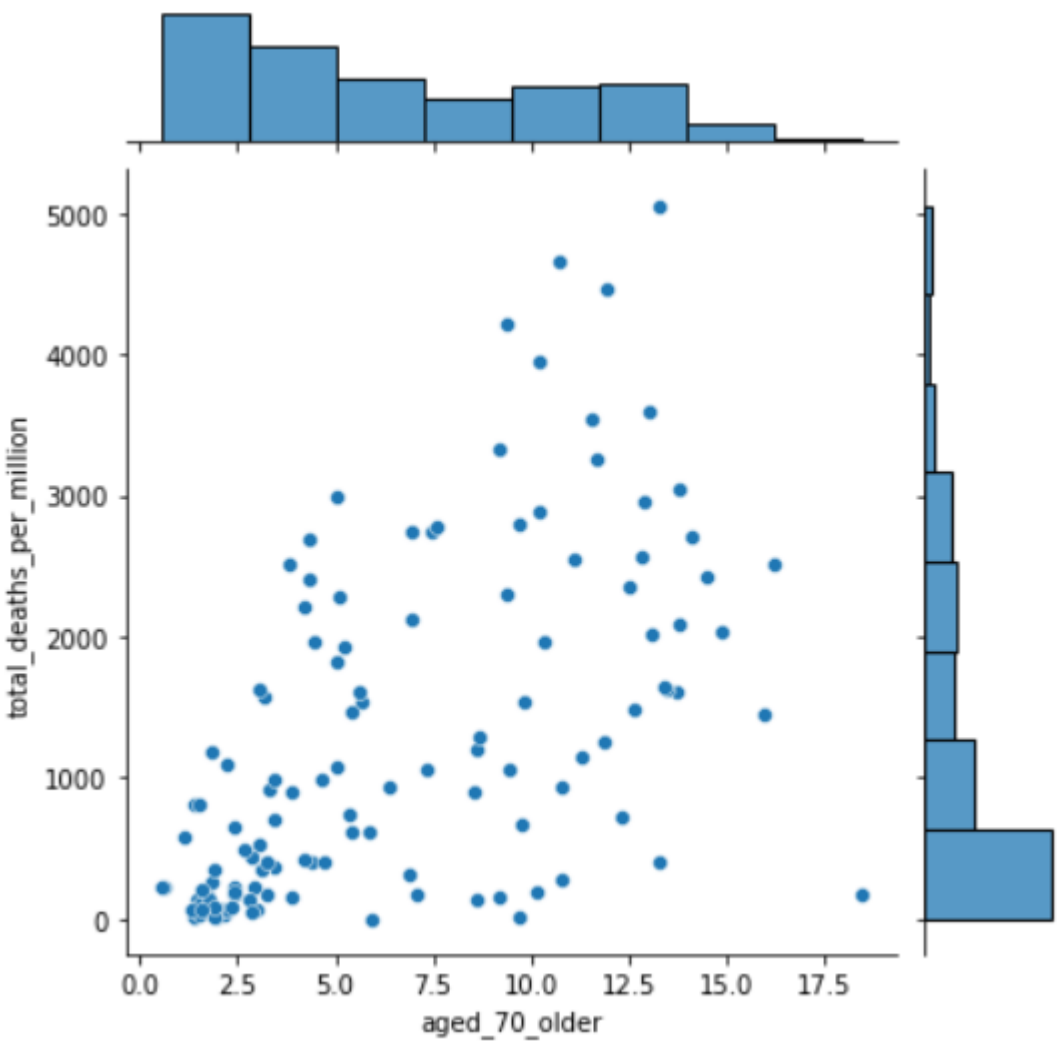
Hex plot using seaborn for the above scatter plot

Hex plot is similar to scatter plot but as the points in the area of hex gets increased the hex will start turning dark. The Hex with less number of scatter points inside is light in colour where as the Hex which has more number of scatter points inside is dark in colour, because of this it is much more quickly understood by viewers who just want a overview.

Similarly, we want to compare the total deaths per million with aged 65 or older and aged 70 or older values for each country. As both values are quantitative it is best to use scatter plot in this situation.



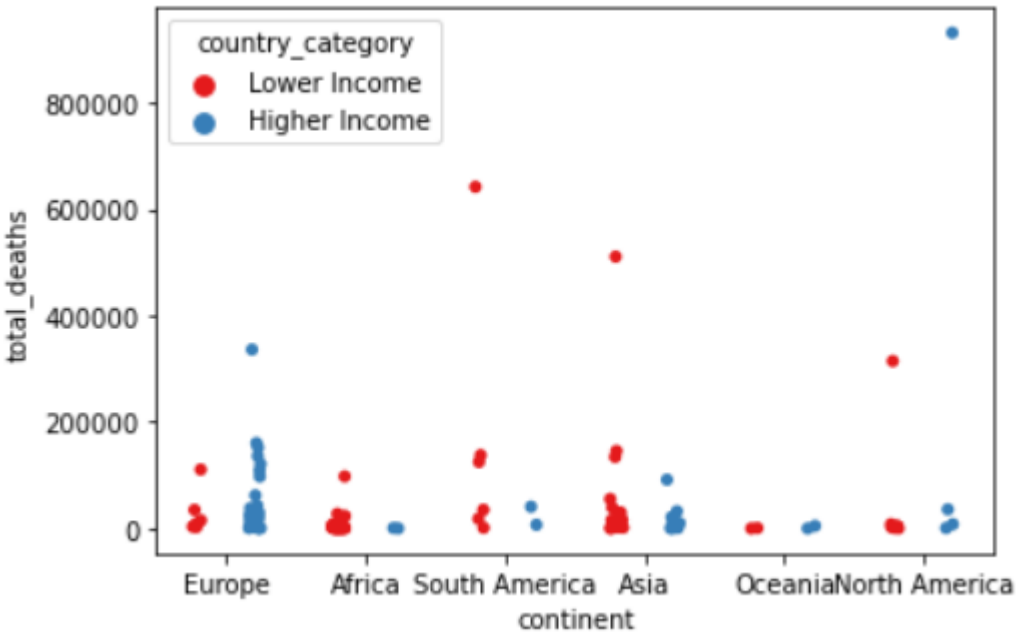
Joint plot between the total deaths per million and aged 65 or older value



Joint plot between the total deaths per million and aged 70 or older value

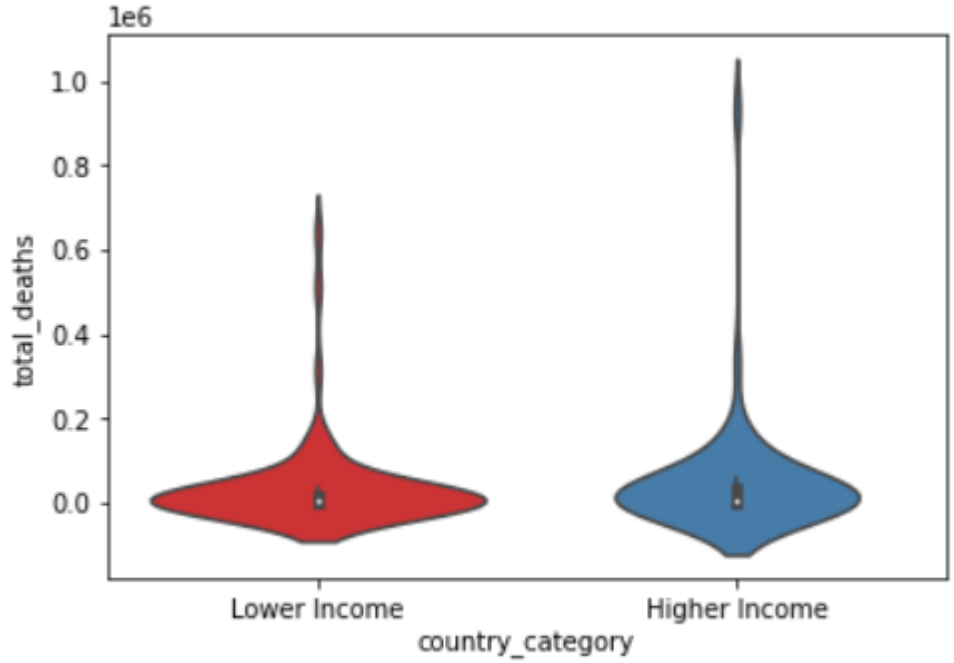
In both the above displayed scatter plot the observations are similar to some extent. We can a cluster at the bottom left side of the graph, less aged population, and low deaths per million. Then as the aged population percentage increases, we can see that the deaths per million values started increasing for most of the points.

Similar to the age factor, we want to analyse the difference between the covid deaths in High income and Low-income countries. We categorized countries based on the GDP per Capita data found in the dataset. There are two categories for the Income status, high income and low income.



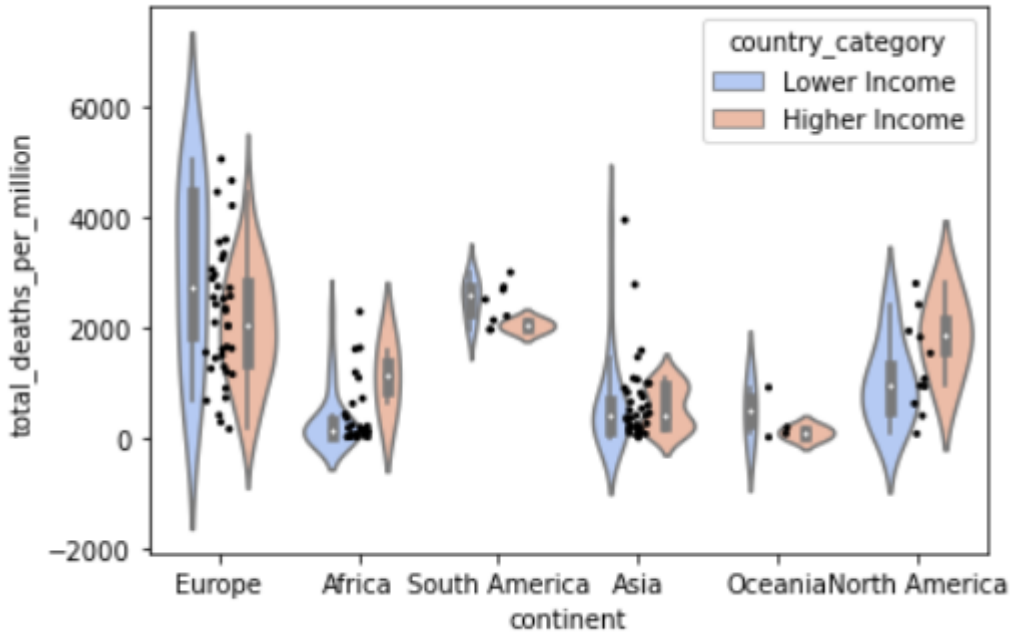
Strip plot for total deaths, Continent and Income status of the country

The above plot is constructed using python library called seaborn. We used this plot as we wanted to plot categorical values with quantitative values, Here we can observe the continent and income status is categorical values and total deaths is quantitative value. We can see the records of the countries in the continent they are present in along with colour based on the income status. Each record in the data frame is plotted on the graph, we can see the outliers in each continent and how many records are present along with their deaths per million and income status.



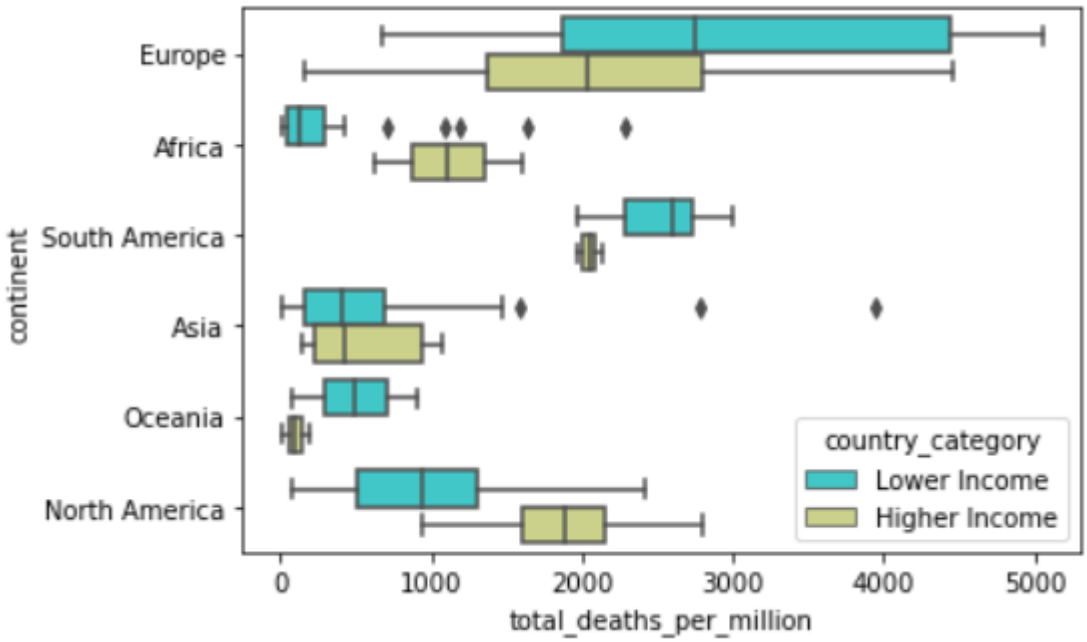
Violin plot is plotted using python libraries

We can notice that the higher income country has higher max deaths per million. The median in lower income is alos less than the higher income. Combining the Violin plot and strip plot we can quickly understand the data as we can see median, quatiles and inner quartile range along with the plots.



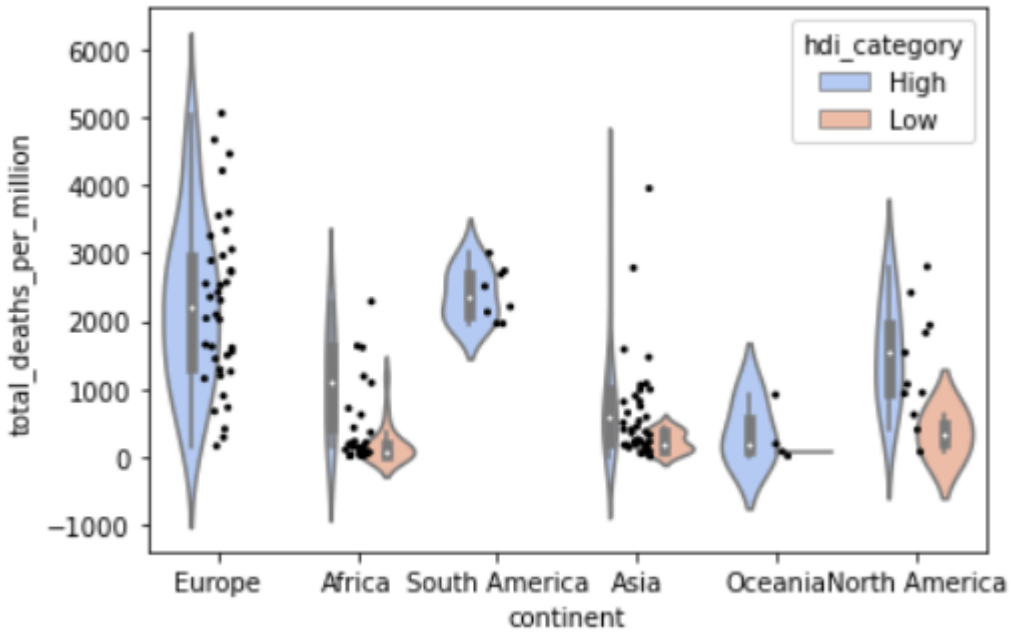
This above plot helps us in understanding as it has both the points and shape of the data frame.

Finally, visuallizing the income status of the country using box plot with continent on Y axis and total deaths per million on x axis and income status of the country as the hue.



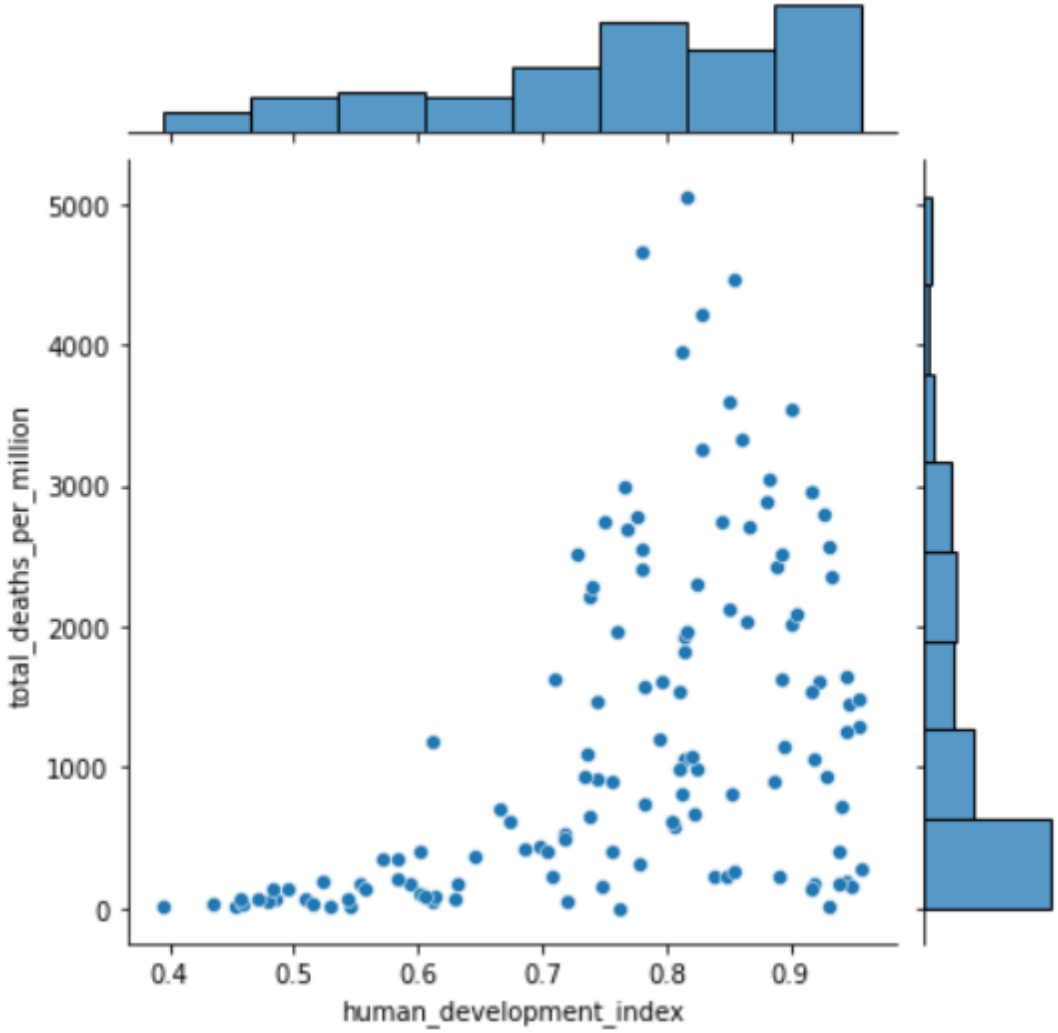
The box plot shows us the distribution of deaths per million in lower and higher income countries in each continent. We can clearly see that Europe continent has significantly higher median than rest of the continents and except in Africa and North America, lower income countries have higher median and max values for Covid deaths per million value compared to higher income countries.

We want to compare human development index value of the country with it's total covid "deaths per million" value. Human development index is one of the widely used measurement in understanding the countries standard. It is generally defined as the average of Standard of living, health, knowledge of the population. We want to check if better HDI countries have less covid deaths compared to lower HDI countries.



Violin and Strip plot for total deaths per million, continent and HDI category

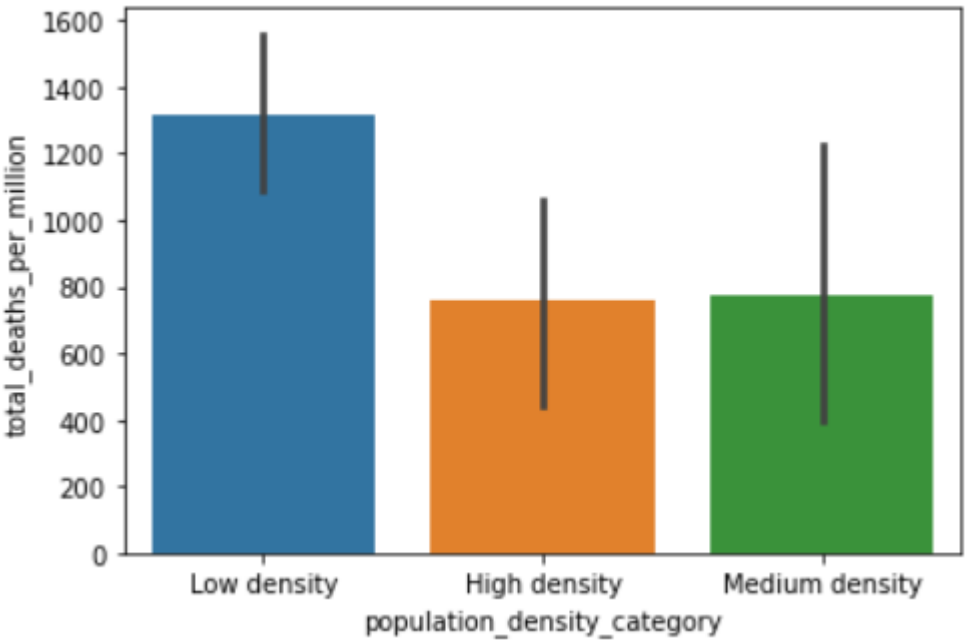
The Combined violin and strip plot is plotted to understand the data distribution in between continents and HDI category. An interesting observation is that there is no low HDI category countries in Europe and Oceania. Europe has the highest recorded deaths per million value even though health, knowledge and standard of living is so high in Europe. On contrary Africa and Asia doesn't have much high HDI countries and most of the countries in these continents is on low HDI zone but they recorded less value per "deaths per million".



Joint plot between “total deaths per million” and “human development index” using seaborn

The above joint plot is plotted between the total deaths per million and human development index. We can observe the distribution of total deaths per million and human development index on the axis and if we look at the scatter plot, we can notice that the higher total deaths per million is present in countries with higher human development index. The possible reason for this is that the countries with higher HDI might have lower population and this might have increased the deaths per million values.

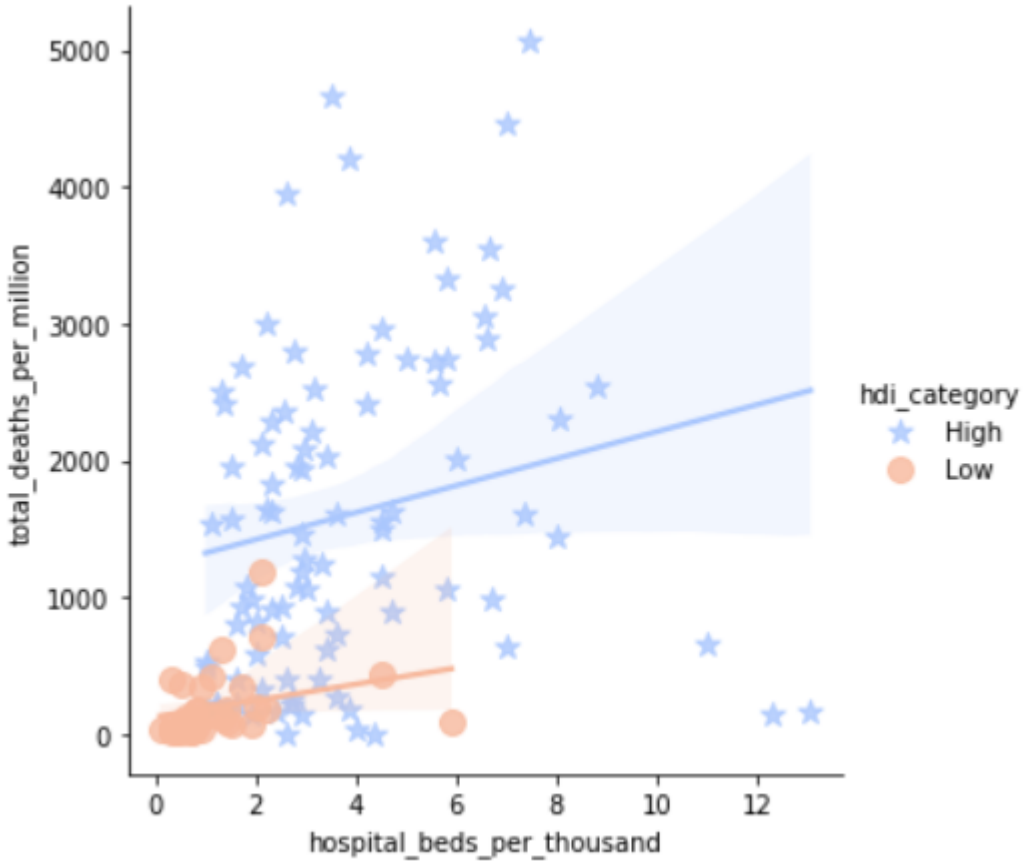
Population density is also widely regarded as an important contributor to the spread of Covid. The population density is categorized into three categories high, medium and low density depending on the population density value in the dataset.



Bar plot on “total deaths per million” and “population density category” created using Seaborn

We can observe from the above displayed image that “total deaths per million” value is higher in lower density region compared to high and medium density. High and medium density have almost similar value.

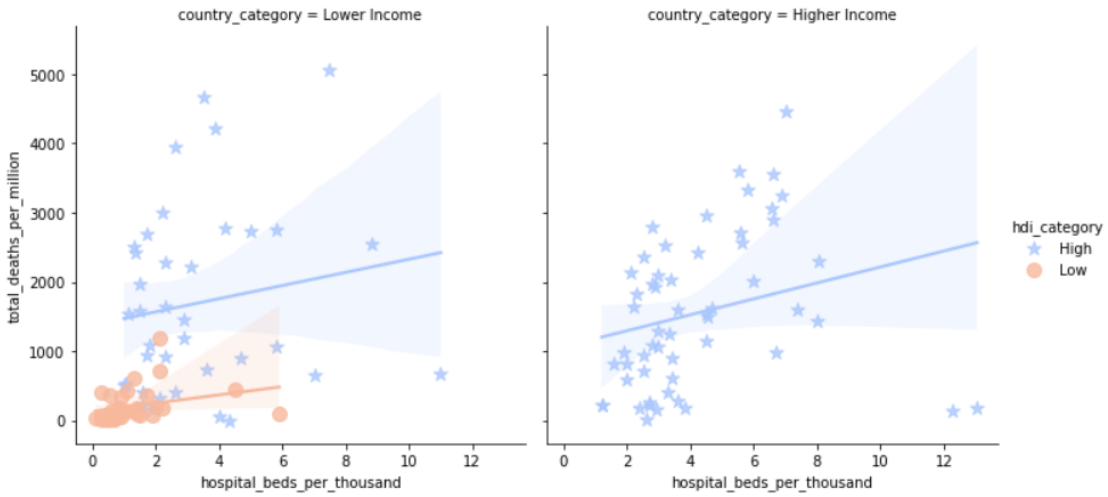
One of the other important parameter to consider is hospital beds for thousand, during the peak of covid, in many countries the beds were not available for seriously affected covid patients. We wanted to explore if the death rate is less in countries which has more hospital beds. The graph is being plotted for hospital beds per thousand and total deaths per million, we used HDI category as HUE and also changed markers for each plot for easy recognition of points category.



Lm plot on “total deaths per million”, “hospital beds per thousand” and “hdi category” constructed using seaborn

From the above displayed graph we can notice that the low HDI category countries has less number of hospital beds for thousand this resulted in forming a cluster of low HDI countries on the bottom left side. Very few of the HDI countries are having significantly higher beds than low HDI countries but the higher HDI countries recorded higher “deaths per million” value.

Now by using Grid functionality we divided the above LM plot into two columns. We passed the country income status as col value. Now the below plot is plotted for four attributes total deaths per million, hospital beds per thousand, HDI category ad Country category.



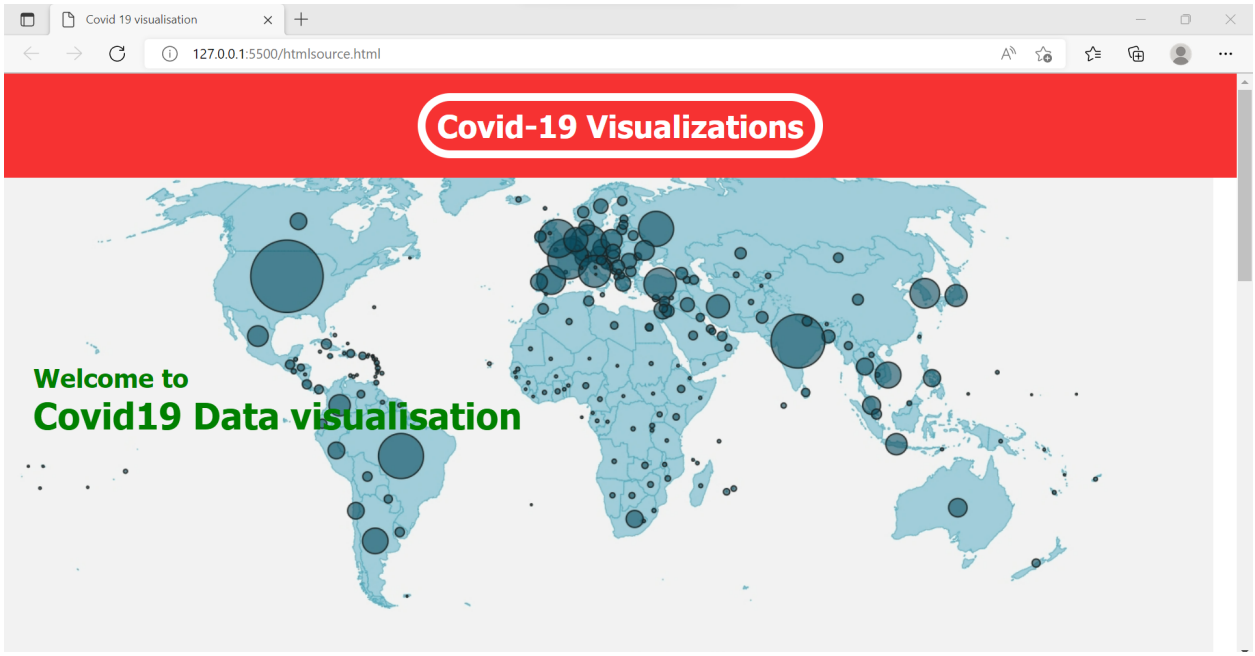
Lm plot for "total deaths per million", "hospital beds per thousand", country category and "hdi category" constructed using seaborn

As we used col value as country category we got two plots, on the left side the plot is for only lower income countries and on the right side the plot is only for higher countries. As hospital beds got increased as we move towards the right side on x axis, so does the deaths per million except in two instances on the higher income plot. We can see two points having near to 12 beds both these points had lower Deaths per thousand values. An interesting observation is that there is no lower HDI countries in Higher income plot.

Implementation of Website using HTML & CSS:

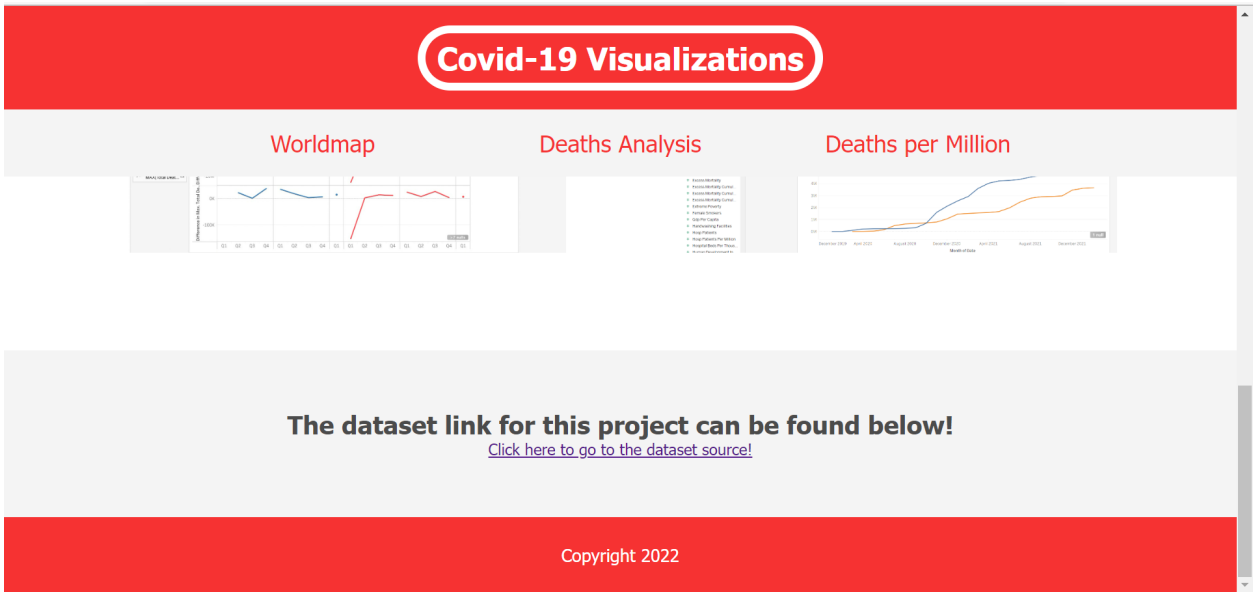
As part of this project, we implemented a website which shows a few important visualizations, an overview of the project can be found on the site. 3 subpages have been created which further show the plots and descriptions for those graphs can also be found there.

Home Page



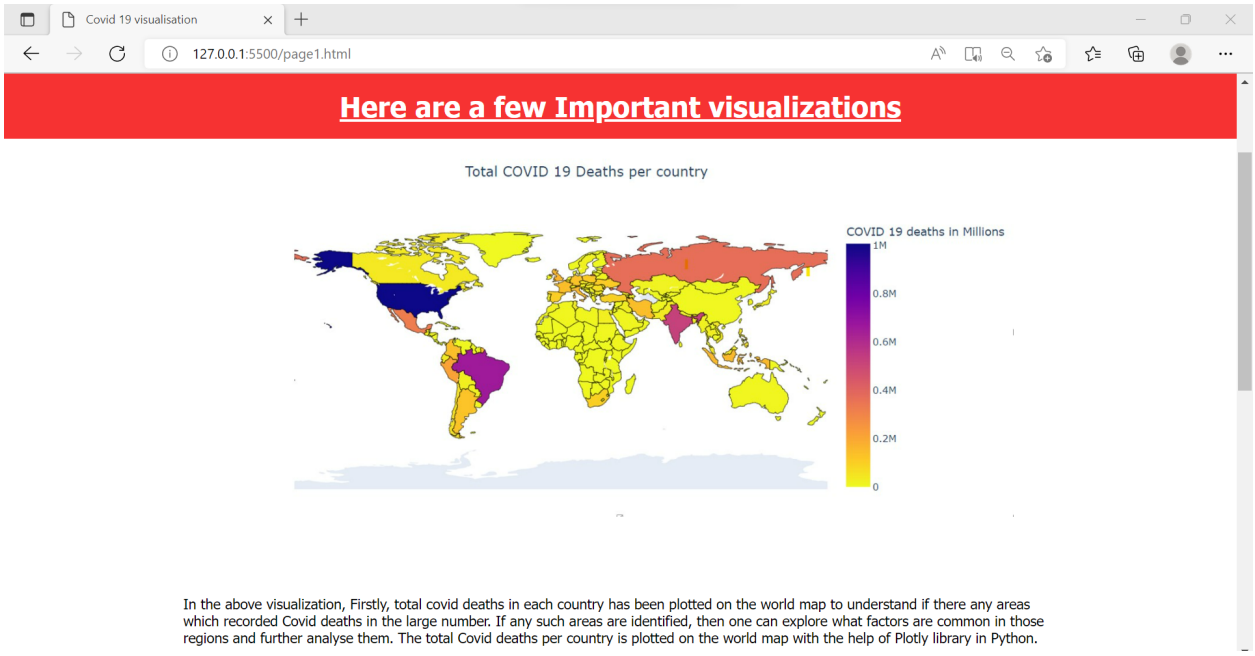
Using various visualisation techniques

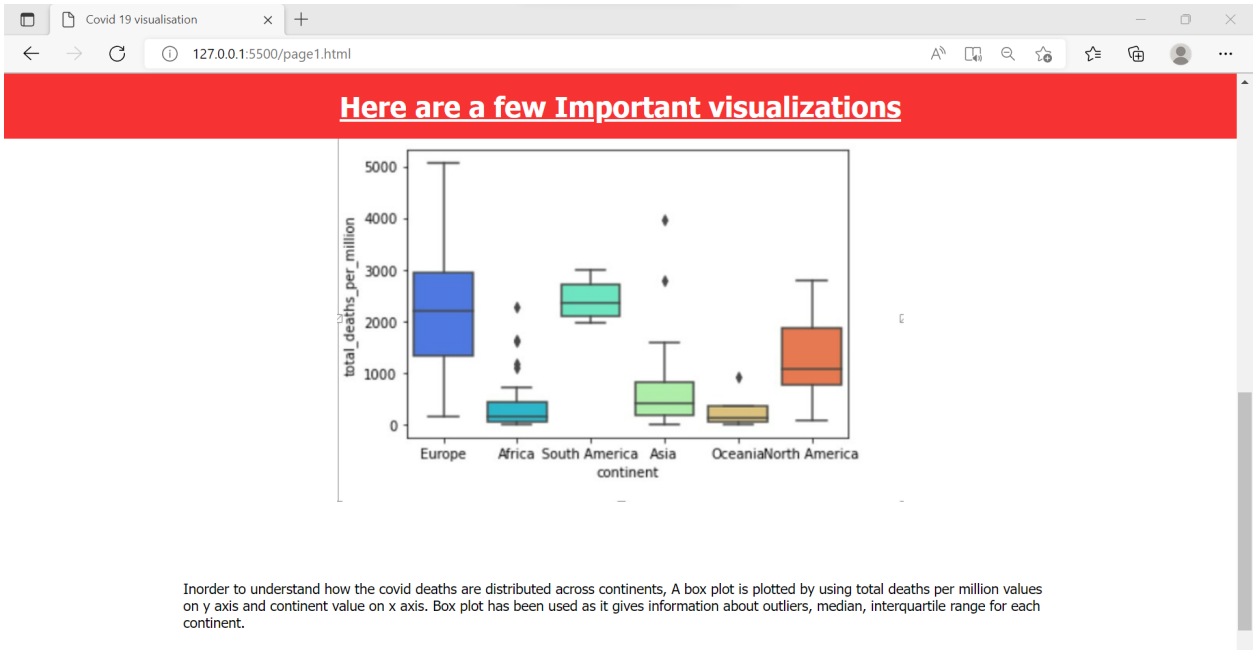
The Covid cases started getting reported from Jan 2020 in Italy, whereas for South Africa the cases started getting reported in March. The cases started increasing in both the countries from September but the reported cases in significantly higher in Italy than South Africa, which is clear from the above line graph. The pattern continued until the end, Italy has 1,23,23,398 by Feb 2022, on the other hand South Africa has total cases of 36,52,024.



The above 3 pics are of the home page. On homepage, we drew the comparison of two countries that have similar project as part of increment1. Those visualizations were put on the home page. Apart from this, you can see 3 sub links namely Worldmap, Death Analysis and Deaths per millions. In those sub-pages, more related visualizations can be found. A link to the source of dataset has been provided, we imported dataset from kaggle as part of this project.

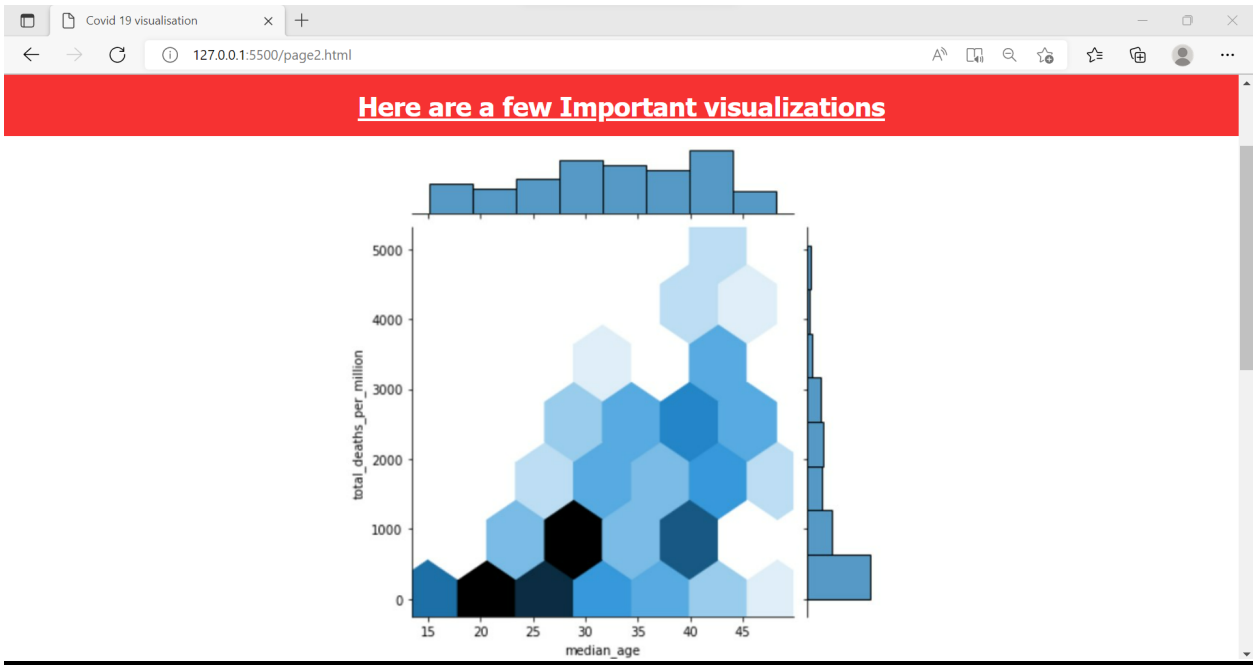
Sub-page1 - Worldmap

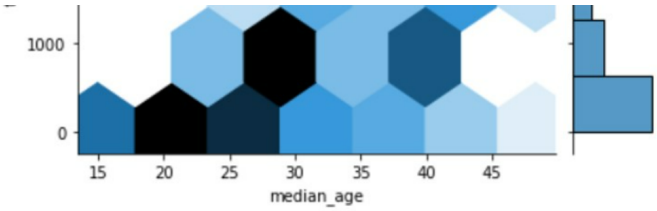




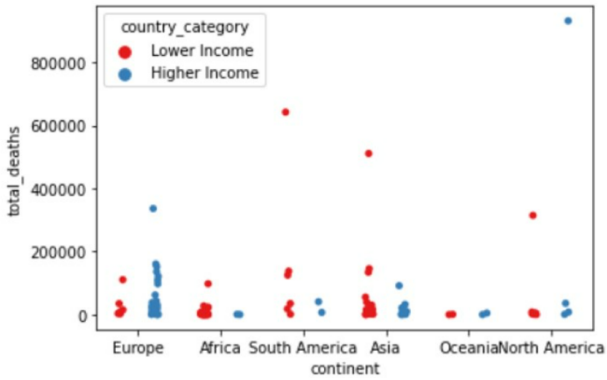
Though we implemented various visualizations, we included the two most important visualizations concerning that area in the website. The same can be found in above pics with description of the visualization. The first one describes about total covid 19 deaths per country with world map. The second one describes about Continent deaths per million averages and ranges.

Sub-page2 – Deaths Analysis:





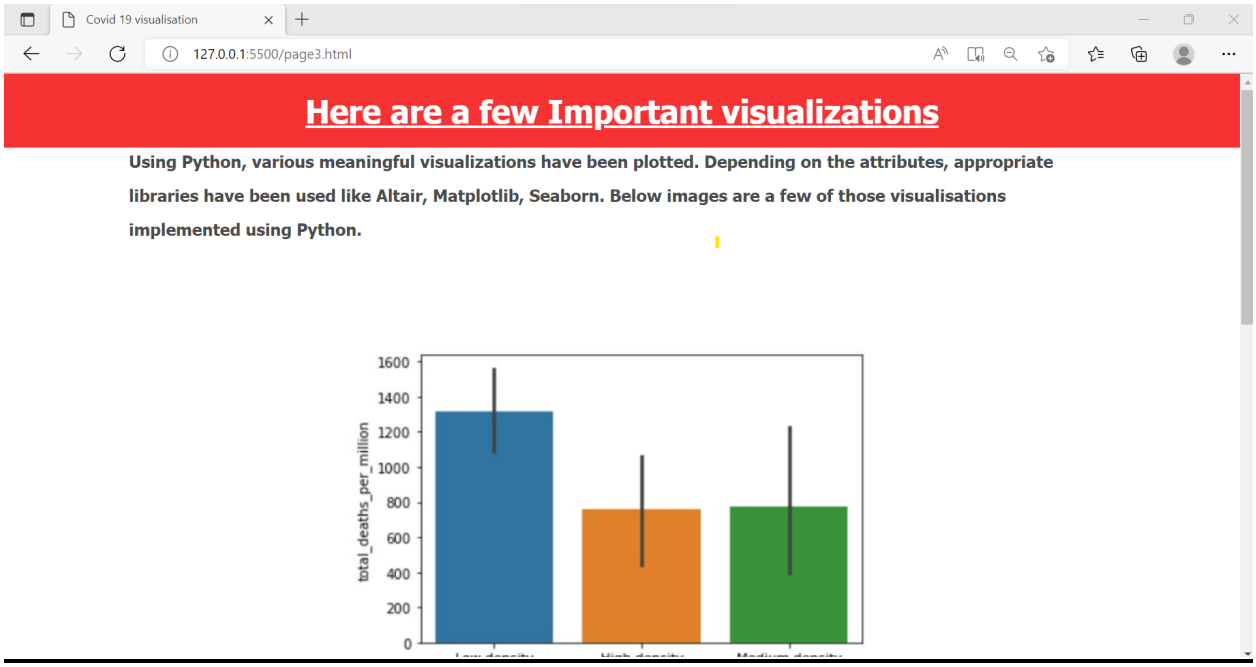
As the points in the area of hex plot increase, the hex will start turn dark. The Hex with less number of scatter points inside is light in color where as the Hex which has more number of scatter points inside is darker in color, because of this it can be quickly understood by viewers who want an overview.

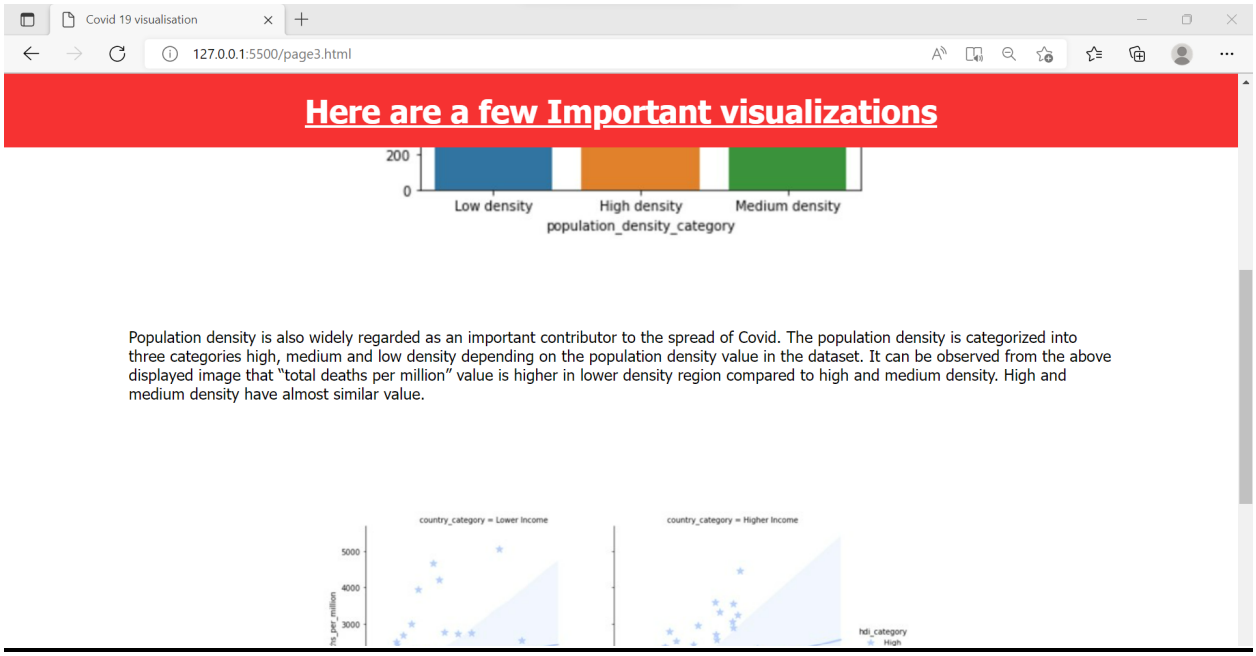


The above plot is constructed using python library called seaborn. We used this plot as we wanted to plot categorical values with quantitative values, Here we can observe the continent and Income status is categorical values and total deaths is quantitative value. We can see the records of the countries in the continent they are present in along with color based on the income status. Each record in the data frame is plotted on the graph, we can see the outliers in each continent and how many records are present along with their deaths per million and income status.

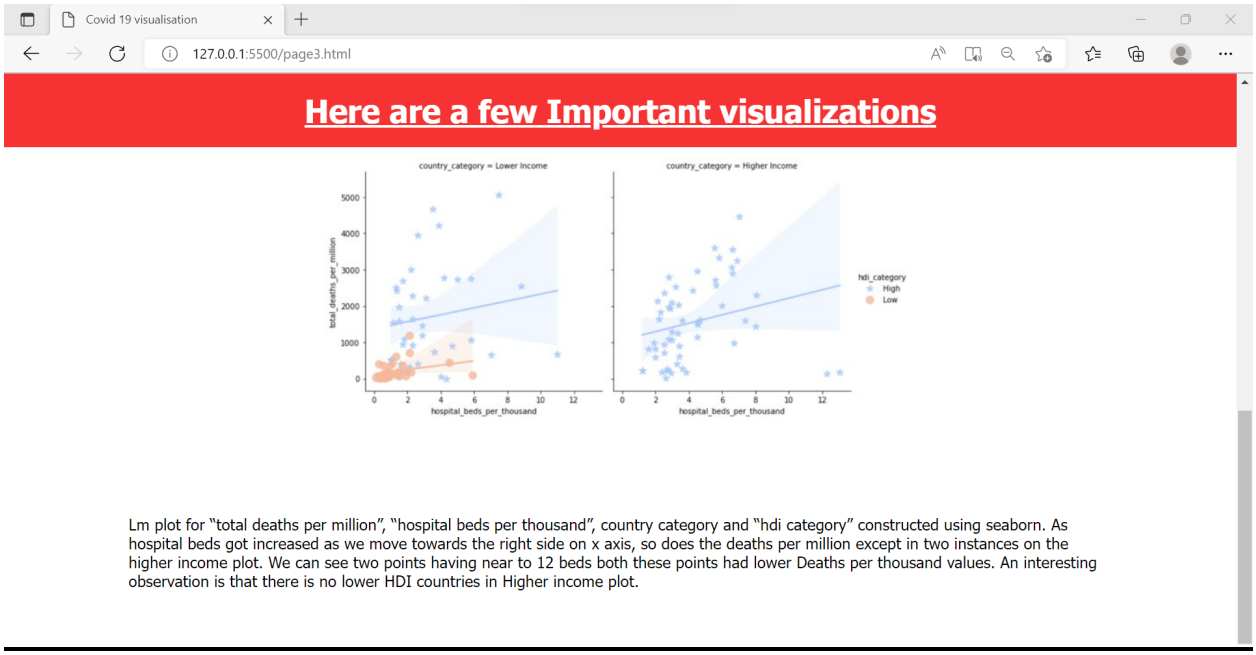
In the second subpage, we included visualizations related to total deaths across continents categorized by lower income and higher income and a hex plot has been plotted between median age and total deaths per million

Subpage-3 Deaths per million





Population density is also widely regarded as an important contributor to the spread of Covid. The population density is categorized into three categories high, medium and low density depending on the population density value in the dataset. It can be observed from the above displayed image that "total deaths per million" value is higher in lower density region compared to high and medium density. High and medium density have almost similar value.



In the 3rd subpage, graphs have been produced that involves the attribute total deaths per million in comparison with other attributes to draw some inferences.

Story Telling:

Chapter-1:

Who

Healthcare analysts, researchers and other communities and stakeholders. Older people, people of particular geographical location, people suffering from chronic illness are in need of help. Business communities were also among the ones that were impacted due to covid.

What

It is a virus called SARS-CoV2 that has caused the pandemic spread all over the world impacting many people and communities directly and indirectly. Older people and people with chronic diseases and underlying conditions have been the most impacted set of people either in the form of death or post covid complications.

When

The first known outbreak started in November 2019. It has been declared a pandemic by the WHO on March 11, 2020. It happened to spread all over the world rapidly later.

Where

1.) This problem possibly occurred in China in 2019 and was spread rapidly around the world now a serious issue all over the world.
2.) There isn't one particular location where this problem doesn't exist right now. It is spread all over the world.

Why

It is said that the root cause/origin of this virus is that it has come through the mammal bats. However, it remains unclear how the virus first spread to humans.

How

Covid-19 was first assumed to have infected humans at an open-air wet market in Wuhan, China. Later theories speculated that it may have been developed as a biological weapon in a Chinese laboratory. Later this was spread from one individual to other individual.

Chapter-2:

Who

The dataset is about covid cases and information of covid in various countries. The dataset has several factors that pertain to covid-19. For particular attributes in some countries a few records are missing due to factors like information available locally. This is an open source dataset that the author has been collecting data from various other source.

What

From the attributes that were available in the dataset such as population, death percentage, age_65_older, GDP_per_capita, median_age and so on few observations can be drawn and the behaviour of covid cases, deaths can be found. It targets the above activities. One can successfully predict some behaviours using the vast information set that is available.

When

This event first took place in 2019 and since then the observations and behavior of the pandemic has been continuously evolving due to various factors. The information provided in this dataset keeps getting updated every week and is a real time data. There are both historic and latest data available in this dataset. The countries that has similar count of population, and other factors like GDP, human_development_index and median_age can see almost the same behavior but only limited to few countries and can't be generalized to every country.

Where

This is a global dataset as it involves all the countries. The data is collected from various grass root levels and is consolidated into one big dataset. it spans across different geographies. The coverage looks different from region to region and is independent of one another.

Why

The aim was to help the communities and the people that were affected by providing sufficient precautionary measures. For example upon looking at the raise in cases and death rates of a particular region or age group, government can improve the vaccination drives throughout the region.

Chapter-3:

Who

The main target is the general public, health analysts, government, anyone that wish to access any information/comparisons on covid related data and that wants to make use of analysis for application in real world.

What

The visualizations provide the general idea on the factors that contribute to the increase in covid cases and deaths. The visualizations show the global effect and the rise in cases day to day.

When

The user can use the visualizations to find any patterns in the data and also to understand the reasons that influence covid cases and deaths.

Where

The visualizations can be deployed in websites and media as a news article for the general public to increase awareness.

Why

The visualizations are useful as it helps them understand the severity of the situation and raise awareness.

How

The community can pass the visualizations among them as visualizations portray the harsh truth better than the plain numbers. They can find motivation to find necessary steps to fight the battle against the covid more affectively.

Project Management Implementation Status Report Work completed:

Implementation status report:

Work completed:

Description:

The visualizations have been done using Tableau, Python and HTML and CSS. We have analyzed the possible contributors for the increase of cases and deaths across countries. The story telling part, documentation and created a web page where important visualizations have been included with analysis.

Responsibility (Task, Person):

SL.NO	Task(s)	Member Name
1	Analysis & Visualizations using Python	Likith Guduru, Sai Sandeep Gollamudi
2	Analysis & visualization using Tableau	Dinesh Tadepalli, Sai Sandeep Gollamudi, Likith Guduru
3	Web page using HTML, CSS	Dinesh Tadepalli
4	Documentation	Naveen Paritala, Dinesh Tadepalli, Sai Sandeep Gollamudi, Likith Guduru

Contributions:

SL.NO	Member Name	Percentage
1	Likith Guduru	100%
2	Sai Sandeep Gollamudi	100%
3	Dinesh Tadepalli	100%
4	Naveen Paritala	100%

References:

Kaggle link for dataset:

https://www.kaggle.com/georgesaavedra/covid19-dataset

Reference Visualizations:

https://www.tableau.com/covid-19-coronavirus-data-resources