**Introduction:**

****

Performing the Exploratory Data Analysis on Covid-19 data set. Using the real time dataset from <https://ourworldindata.org/covid-deaths>.

**Examining the data:**

The data downloaded is in .csv format which contains 67 columns and 240451 rows.

The dataset contains the information related to covid cases in all countries with their corresponding population, count of covid cases, deaths, vaccinations incorporated and all.

As the dataset has too many attributes, to gauge the data we are splitting the dataset into two files as CovidCasesnDeaths.xlsx and CovidTestsnVaccines. xlsx.

**Loading the data into SQL server database:**

Created a new Database name Covid\_DB in Microsoft SQL Server Management Studio.

Imported both the .xlsx files into SQL server using the SQL Server Import and Export Wizard.

The data from the files are saved as tables in Covid\_DB as CovidCasesnDeaths and CovidTestsnVaccines. On observing the column names and their data types, it is found that few attributes which are count based are configured as float instead of int/bigint. So, while querying those columns relevant datatype are assigned.

**Exploratory Data analysis:**

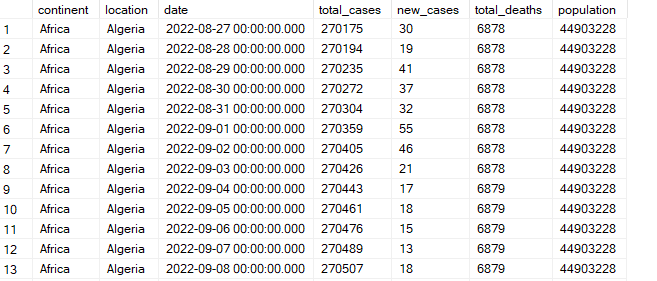
**1. Fetching the basic details data of only countries:**

select continent, location, date, total\_cases, new\_cases, total\_deaths, population

from Covid\_DB..CovidCasesnDeaths

where continent is not null

order by 1,2



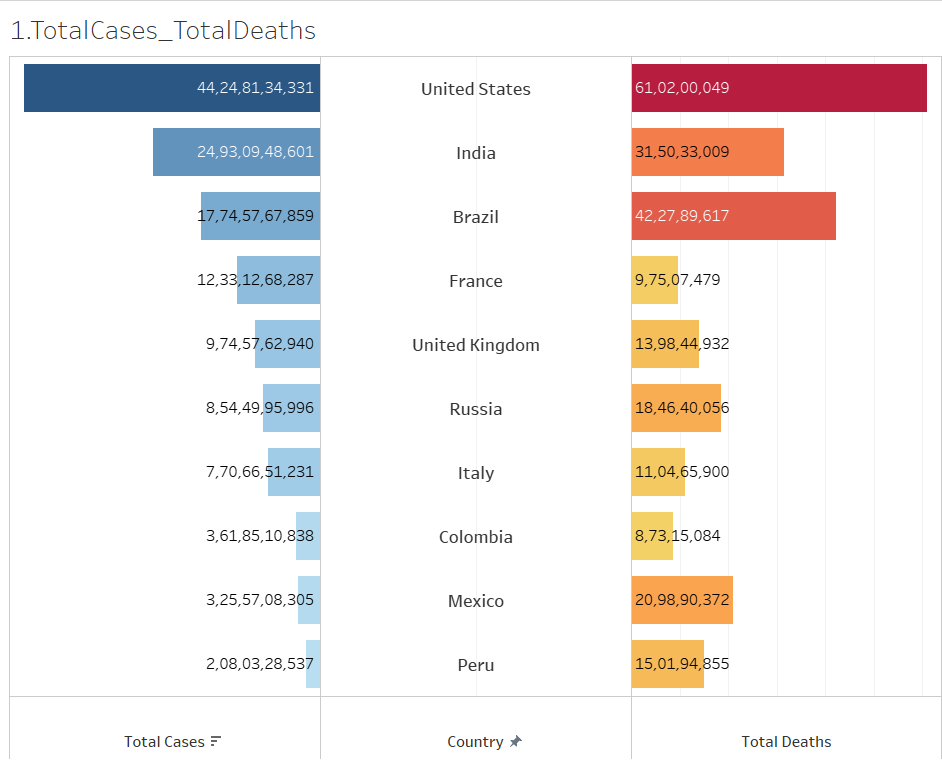
*Key findings:*

Gives the brief view of data for each country, cases and deaths registered on daily bases.

Below are the visualizations made using the Tableau Desktop Trial Version.

***.*** *Comparing the Top 10 deaths of countries by total cases.*

From the butterfly bar chart, we can notice that United States has more total cases and even the death counts.

****

***.*** *Average deaths by months:*

When we look into average deaths by months continents wise, it can be observed that North America has more deaths during the months April to December. Where in continent with least population shows less average deaths.

****

**2. Finding the death percentage over total cases registered in country India:**

select location, date, total\_cases, total\_deaths,

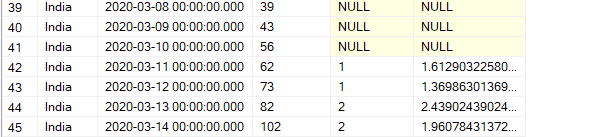
(total\_deaths/total\_cases)\*100 as DeathPercentage

from Covid\_DB..CovidCasesnDeaths

where location = 'India'

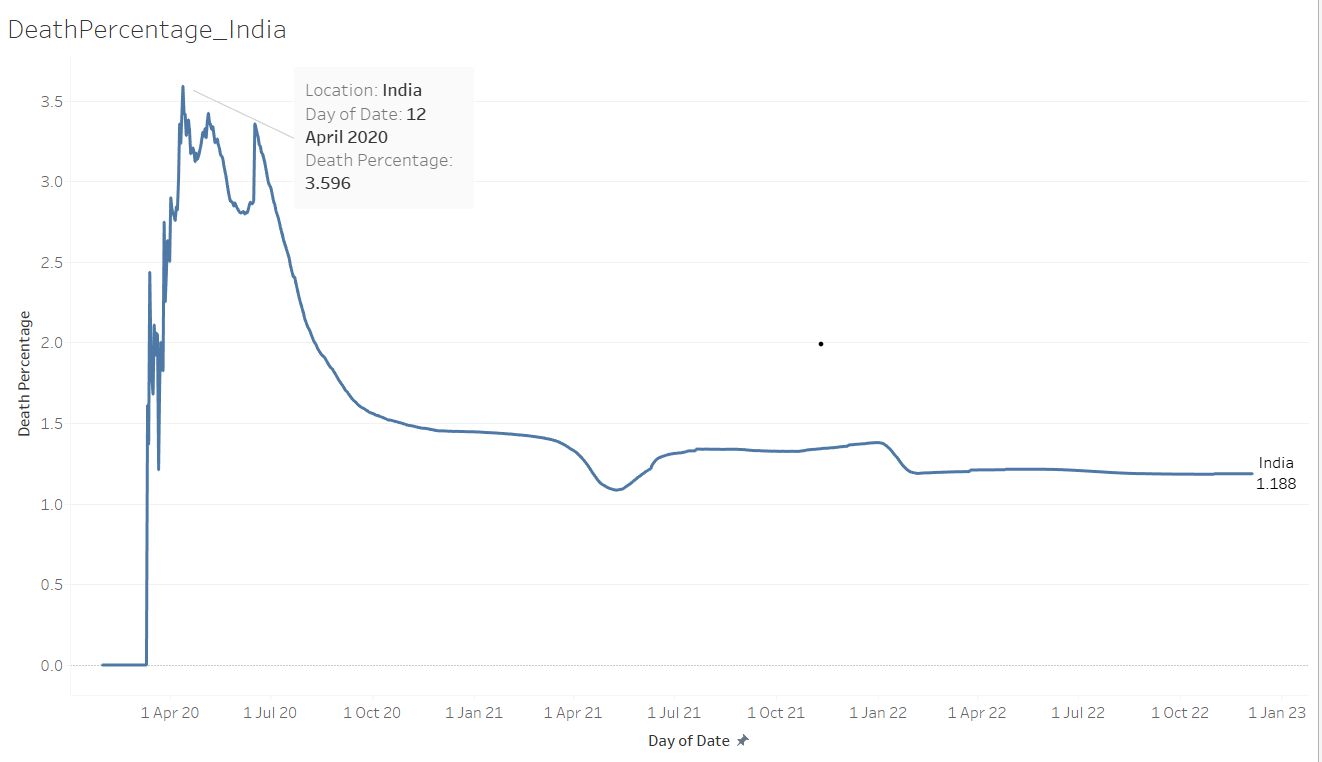
and continent is not null

order by 2

****

*Key findings:*

When considered India as a sample to observe the total death percentage over total cases registered over the period, the percentage is more during the beginning year of the pandemic. We can understand result set data better by finding the trend with the help of visualization further.



A line chart is visualized using the parameters death percentage and time period, here we can understand that India has seen the peak of death 3.596% during April 2020. But since October 2020 death rate was under 1.5%.

But now let’s try out the **death percentage by grouping according to the severity**.

with temp\_tb as

(select location, date, total\_cases, total\_deaths,

(total\_deaths/total\_cases)\*100 as DeathPercentage

from Covid\_DB..CovidCasesnDeaths

where location = 'India' and continent is not null

--order by 2

)

select

year(date) as date\_year,

SUM(case when DeathPercentage >= 3 then 1 ELSE 0 END) as 'High',

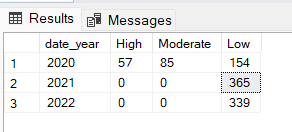
SUM(case when DeathPercentage >= 2 and DeathPercentage < 3 then 1 ELSE 0 END) as 'Moderate',

SUM(case when DeathPercentage < 2 then 1 ELSE 0 END) as'Low'

from temp\_tb

group by year(date)

order by 1

****

*Key findings:*

We can notice that even though in the beginning year 2020 the severity is varied from high to low, but over the next consecutives years the death percentage remained consistently low.

**3. Countries with infection Rate (total cases) compared to population:**

select location, population, MAX(total\_cases) as HighestInfectionCount,

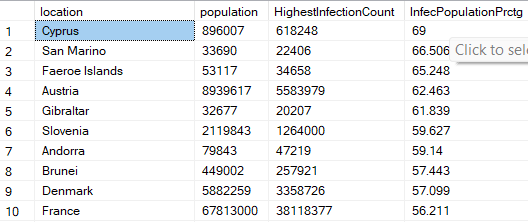
ROUND(MAX((total\_cases/population))\*100, 3) as InfecPopulationPrctg

from Covid\_DB..CovidCasesnDeaths

where continent is not null

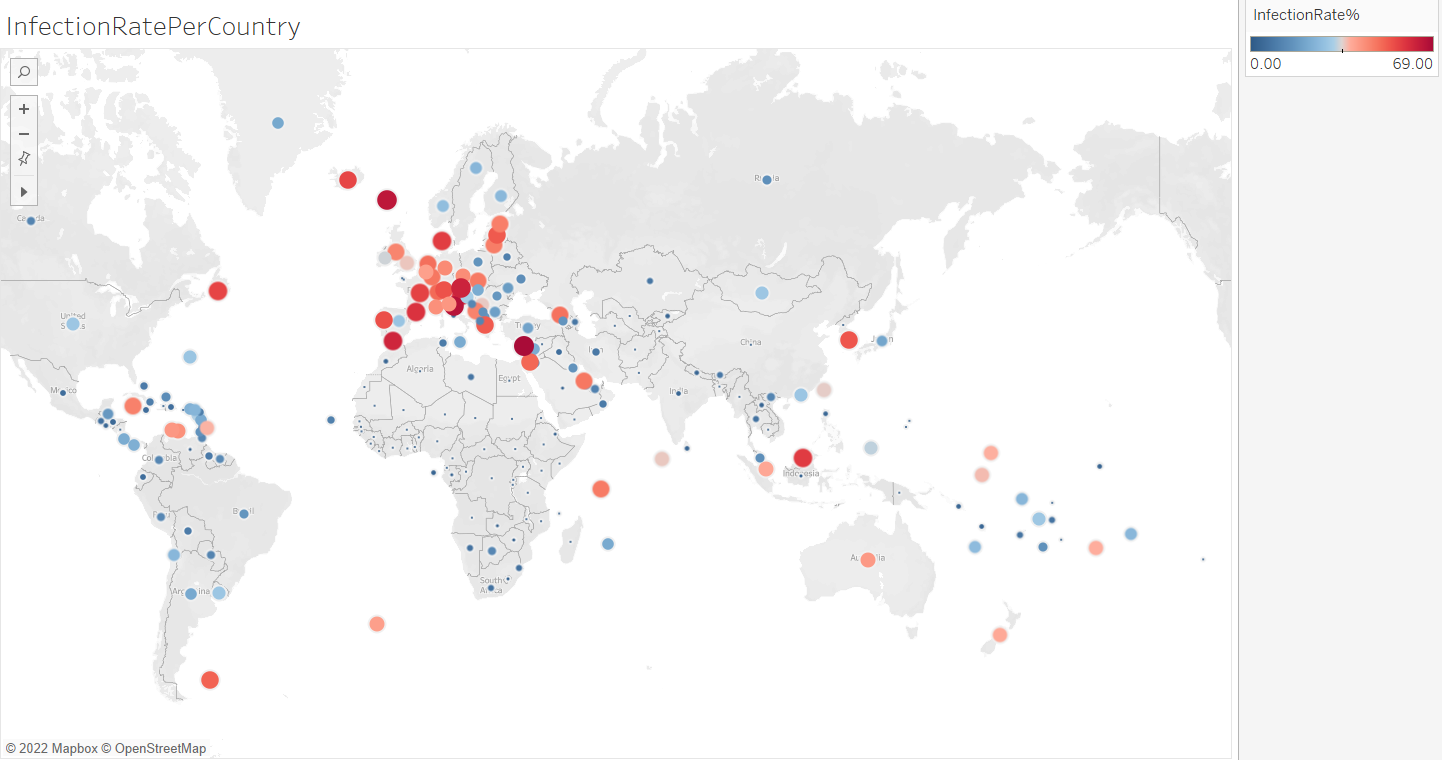
group by location, population

order by 4 DESC



*Key findings:*

Cyprus country shows the highest infection of 69%, though it has low population. China has lowest percentage of 0.12 as it’s the highly populated country. We can get more insights by visualizing this result set.



Using the point map, red colour data points indicates high death rates. The high death rates are concentrated at eastern region of European continent.

**4. Countries with death count per population:**

select location, MAX(population) as CountryPopulation

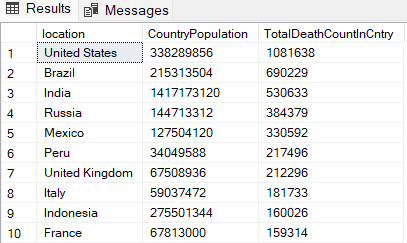
,MAX(cast(total\_deaths as int)) as TotalDeathCountInCntry

from Covid\_DB..CovidCasesnDeaths

where continent is not null

group by location

order by 3 DESC

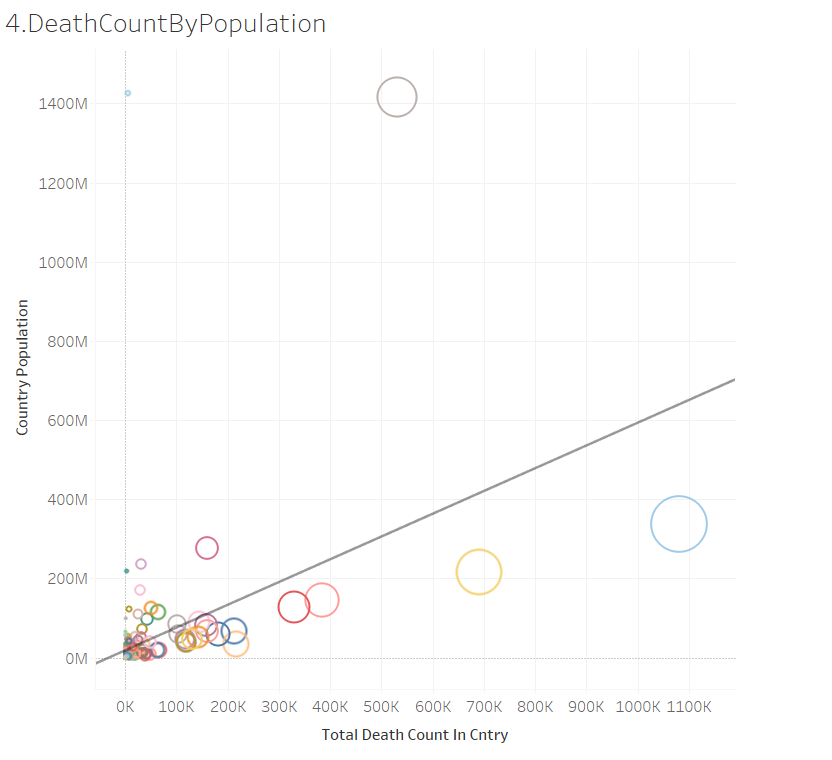


*Key findings:*

United States shows the highest death toll 1081638. Countries like Saint Pierre and Miquelon,

Cook Islands shows minimum death count of 1.

We can understand the correlation between the attributes population and death count using the scatter plot during viz.



**5. Average infection rate of countries by population:**

select location, max(population) as CountryPopulation,

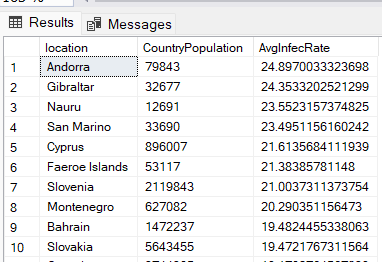
AVG((cast(total\_cases as int)/population)\*100) as AvgInfecRate

from Covid\_DB..CovidCasesnDeaths

where continent is not null

group by location

order by AvgInfecRateDESC



*Key findings:*

Countries with low population like Andorra, Gibraltar, Nauru, San Marino show high Average infection rate above 20%.

Now let’s check **continent wise Average Infection rate**:

select continent, max(population) as CountryPopulation,

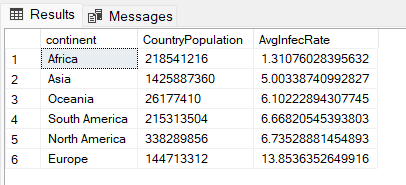
AVG((cast(total\_cases as int)/population)\*100) as AvgInfecRate

from Covid\_DB..CovidCasesnDeaths

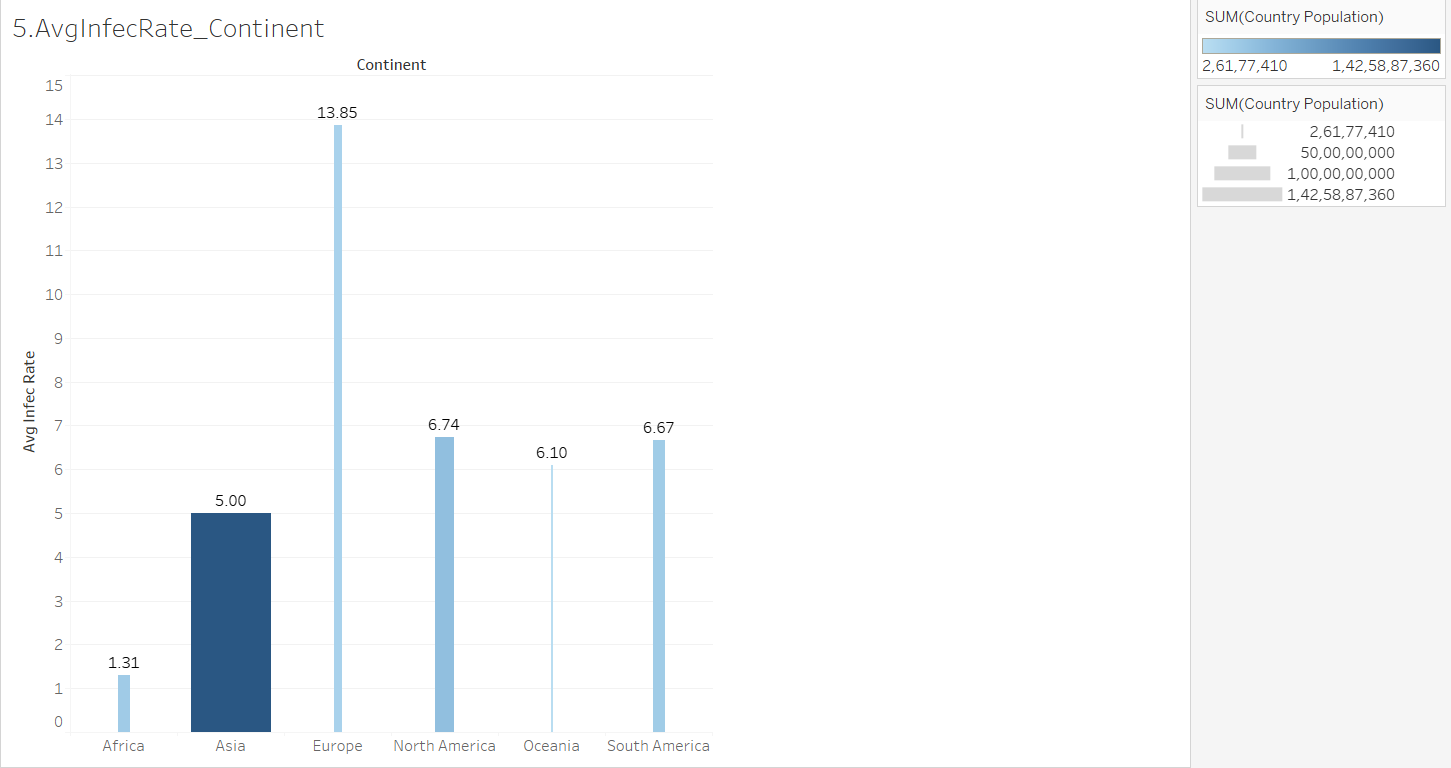
where continent is not null

group by continent

order by AvgInfecRate



*Key findings:* Europe continent shows high average infection rate 13.8%. And Africa shows low average infection rate 1.31%.



Asia with high population shows average infection rate of 5.00%, wherein Oceania with very less population has 6.10% infection rate.

**6. Continent wise covid deaths:**

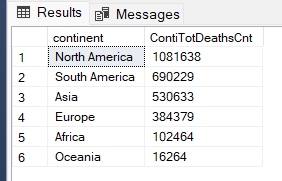
select continent, Max(cast(total\_deaths as int)) as ContiTotDeathsCnt

from Covid\_DB..CovidCasesnDeaths

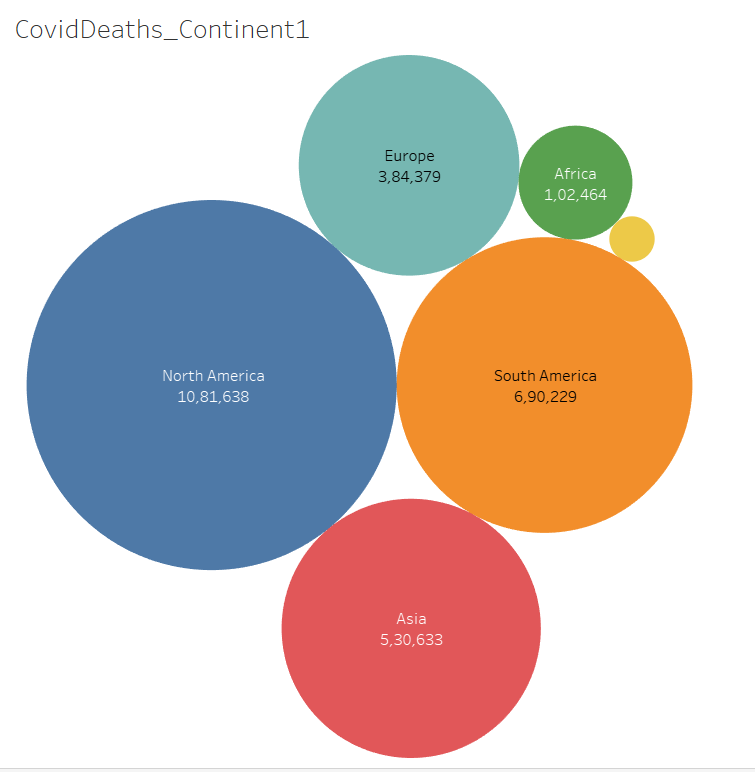
where continent is not null

group by continent

order by ContiTotDeathsCnt DESC



*Key findings:* The resultant data shows that continent North America has highest death toll of 1081638. Oceania has lowest count 16264.

****

**7. Finding the highest average new cases over the year (monthly):**

--select MONTH(date) as monthNo

select DATENAME(MONTH, date) as month

,ROUND(AVG(new\_cases), 0) as AvgNewCases

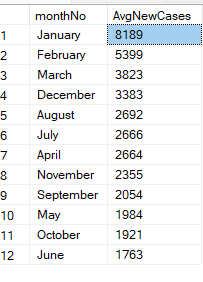
from Covid\_DB..CovidCasesnDeaths

where continent is not null

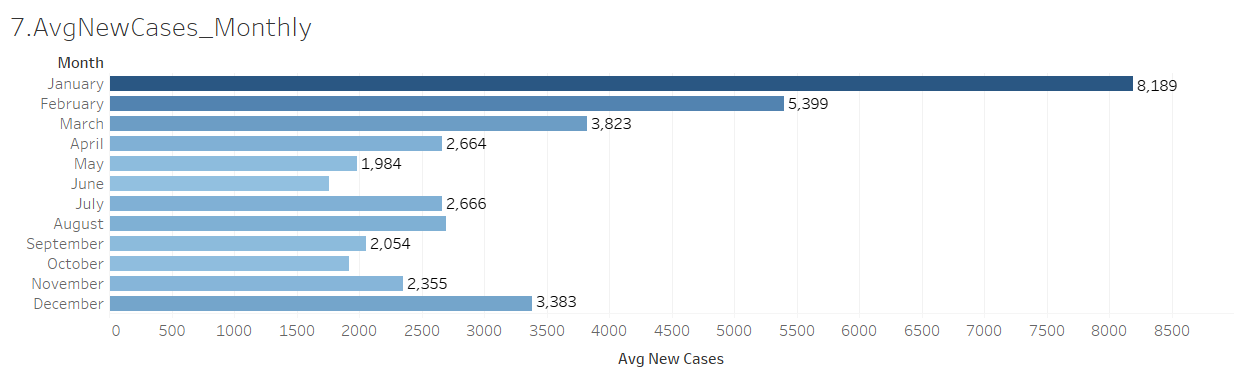
--group by MONTH(date)

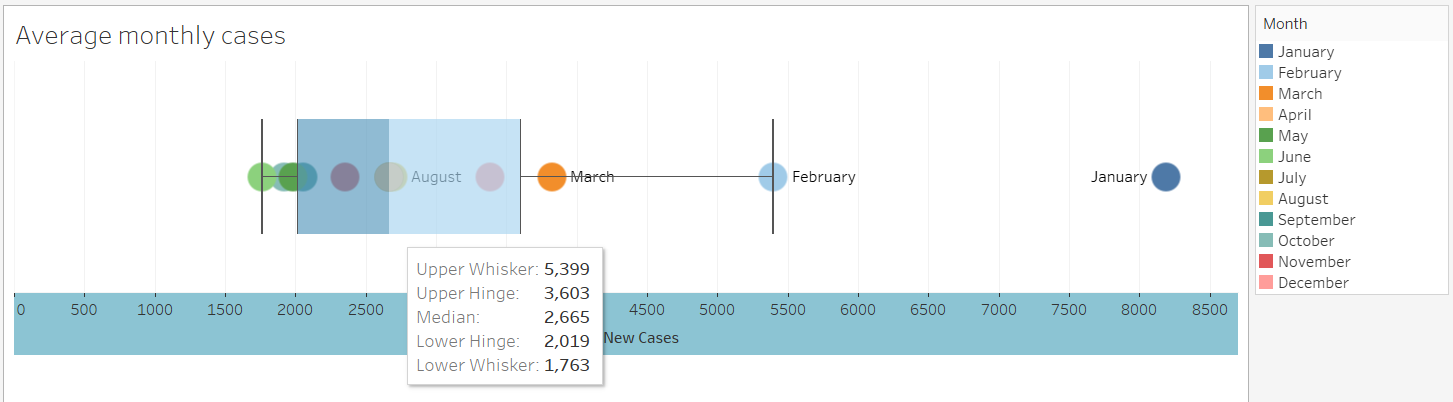
group by DATENAME(MONTH, date)

order by AVG(new\_cases) DESC



*Key findings:* Beginning of the year has high average infection cases and picks up at the end of the year.





**8. Total cases and death percentage in the world:**

select

SUM(new\_cases) as tot\_cases,

SUM(CAST(new\_deaths as int)) as total\_deaths,

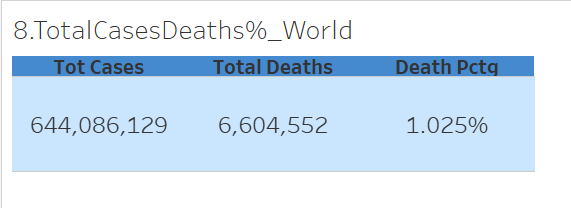
SUM(CAST(new\_deaths as int))/SUM(new\_cases)\*100 as DeathPctg

from Covid\_DB..CovidCasesnDeaths

where continent is not null

order by 1,2

****

****

**9. First case recorded in the countries:**

with temp\_tb as

(select RANK() OVER(PARTITION BY location ORDER BY location, date) as rn,

location, date as firstCaseDate

,total\_cases as firstCasesCount

from Covid\_DB..CovidCasesnDeaths

where total\_cases is not null

and continent is not null

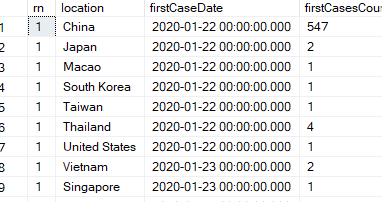
)

select \*

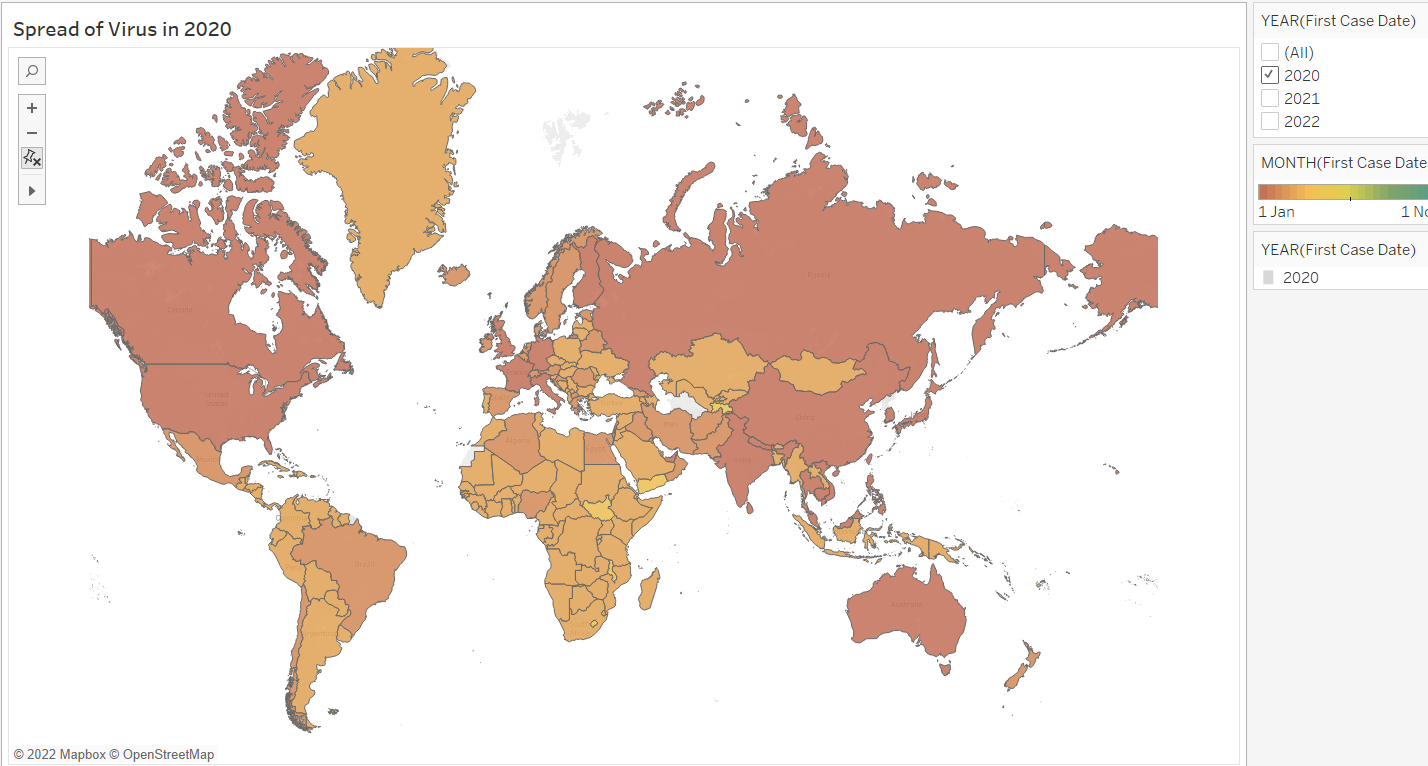
from temp\_tb

where rn = 1

order by firstCaseDate

****

*Key findings:* First case(s) reported in country China on Jan 22nd 2020. Using the map visualization, we can understand how the virus has spread across the globe.

****

When we consider the first case recorded during the year 2020, from the map we can understand that by mid of year 2020 most of the countries across the globe got infected.

Let’s understand the data more by analyzing the tests and vaccinations of countries data along.

**10. Countries with total vaccinations:**

select location,

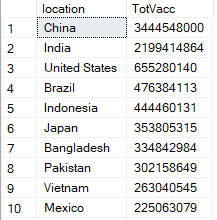
MAX(CAST(total\_vaccinations as bigint))as TotVacc

from Covid\_DB.. CovidTestsnVaccines

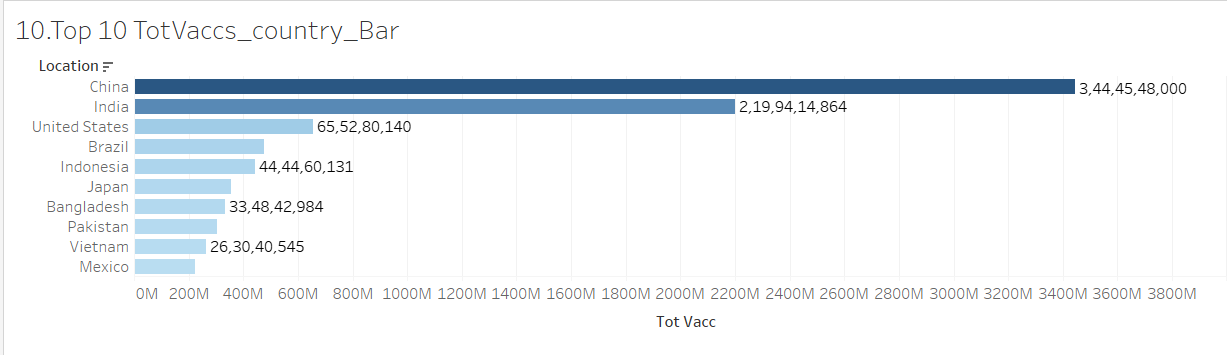
where continent is not null

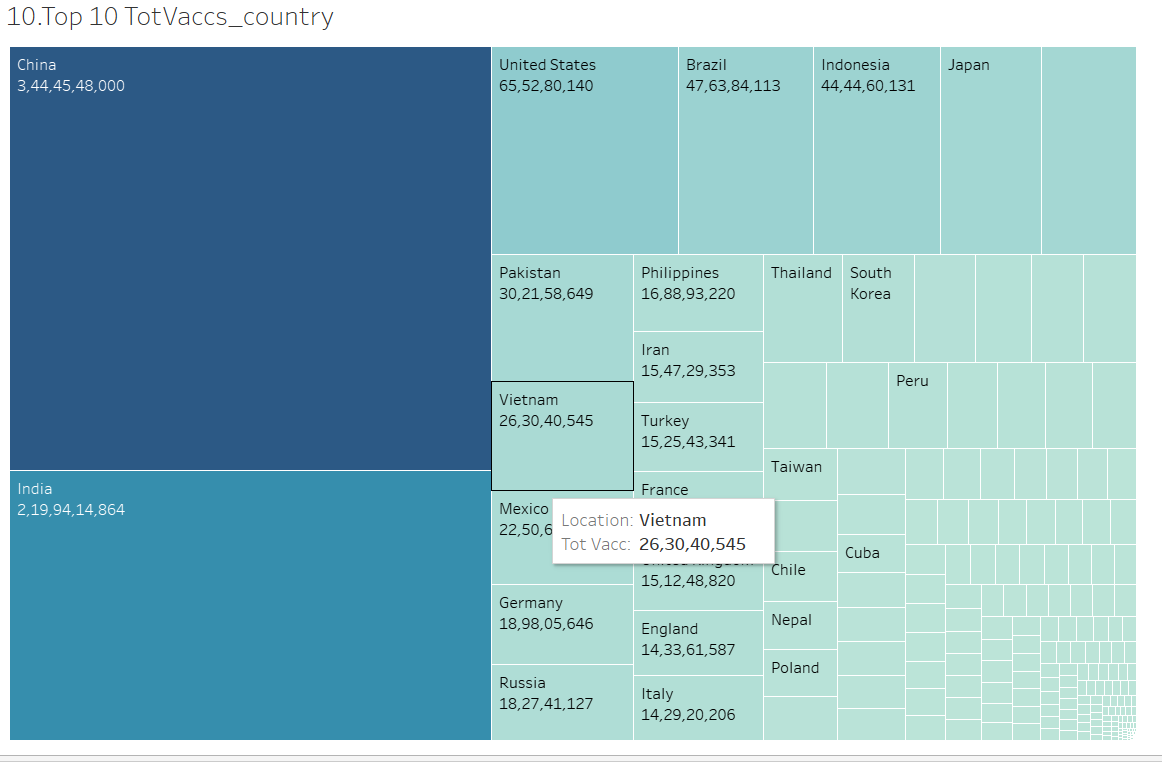
group by location

order by TotVacc desc



Key findings: China has topped in total vaccination count and Pitcairn stands at last.





**11. Finding total vaccinations by population:**

select cd.continent, cd.location, cd.date, population, cv.new\_vaccinations

from Covid\_DB..CovidCasesnDeaths cd

join Covid\_DB.. CovidTestsnVaccines cv

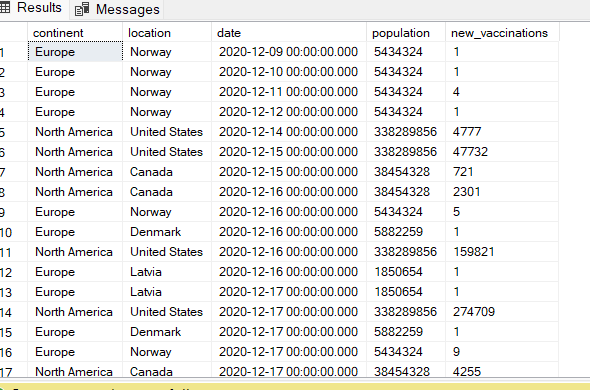
on cd.location = cv.location

and cd.date = cv.date

where new\_vaccinations is not null

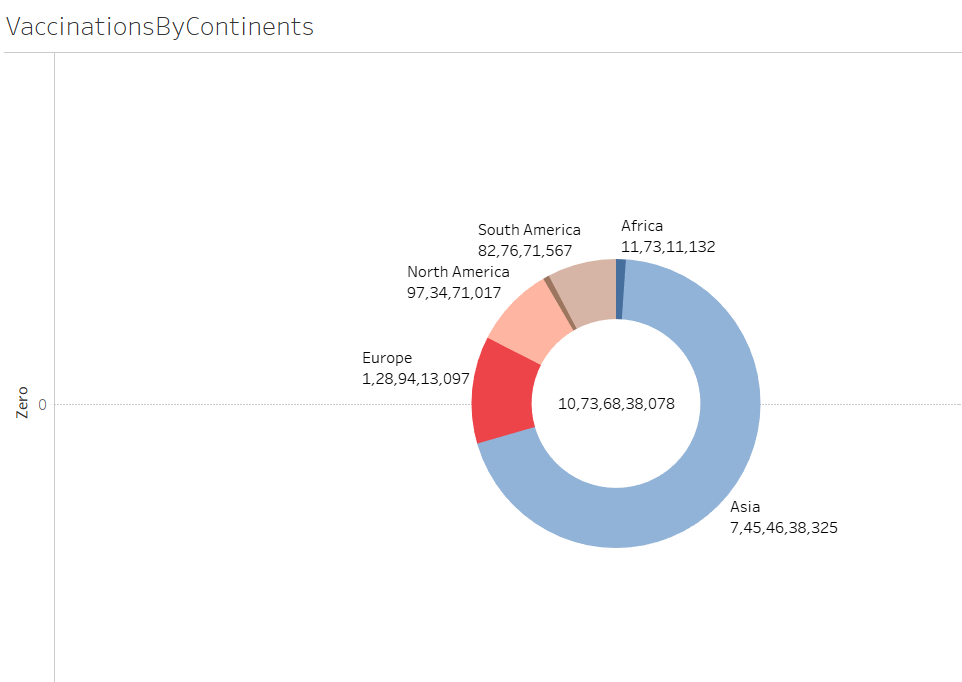
and cd.continent is not null

order by cd.date

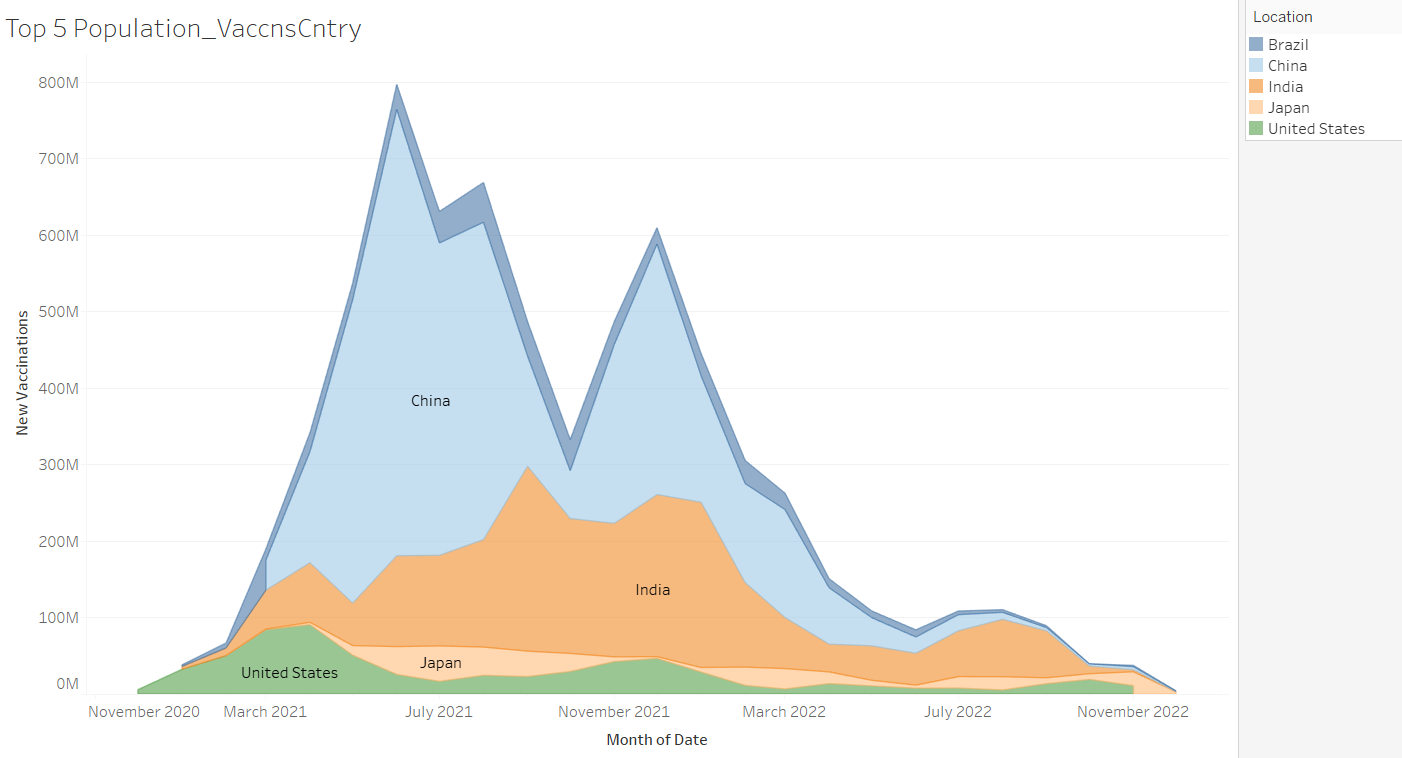


This is the basic result set which can be used for further analysis of different attributes.

Below donut chart shows the vaccinations count in the continents. The chart clearly shows that highly populated continent Asia tops in vaccination count.



Considering the top 5 countries vaccination drive over the period. The chart shows that United states has started the drive earlier in year 2020 where in other countries started in 2021.



**12. Finding the cumulative new vaccinations count:**

select cd.continent, cd.location,cd.date, cd.population, cv.new\_vaccinations,

SUM(cast(cv.new\_vaccinations as bigint))

OVER(partition by cd.location order by cd.location, cd.date) as CumulativeNewVaccCnt

from Covid\_DB..CovidCasesnDeaths cd

join Covid\_DB.. CovidTestsnVaccines cv

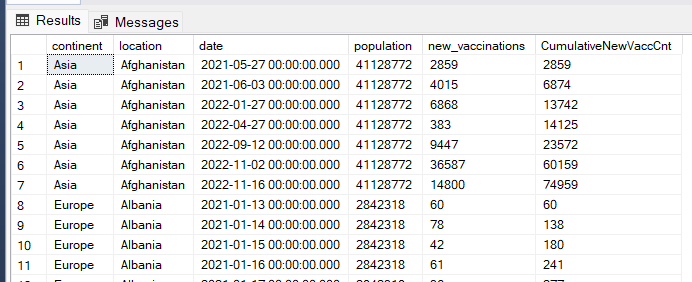
on cd.location = cv.location

and cd.date = cv.date

where cv.new\_vaccinations is not null

and cd.continent is not null

order by cd.location



**Creating a View**:

A view is created from both the join tables:

CREATE VIEW view\_covid\_data as

select

cd.continent as continent,

cd.location as country,

cd.date as rec\_date,

cd.population as cntry\_pop,

cd.new\_cases new\_cases,

cv.new\_vaccinations new\_vacc,

cv.total\_vaccinations as tot\_vacc,

cv.new\_tests as new\_test,

cv.people\_vaccinated as peo\_vacc,

cv.life\_expectancy as life\_expcy,

cv.population\_density as pop\_den

from Covid\_DB..CovidCasesnDeaths cd

join Covid\_DB..CovidTestsnVaccines cv

on cd.location = cv.location

and cd.date = cv.date

where cd.continent is not null

We are going to use this view for our further queries and even for visualization. Let’s continue with the scenarios related to vaccinations

**13. Countries with percentage of population vaccinated:**

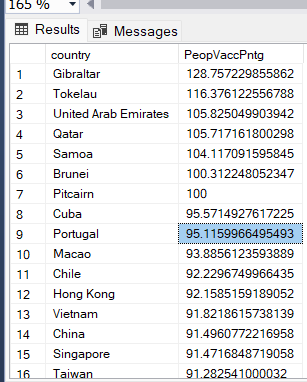
select country,

(MAX(CAST(peo\_vacc as bigint))/MAX(cntry\_pop))\*100 as PeopVaccPntg

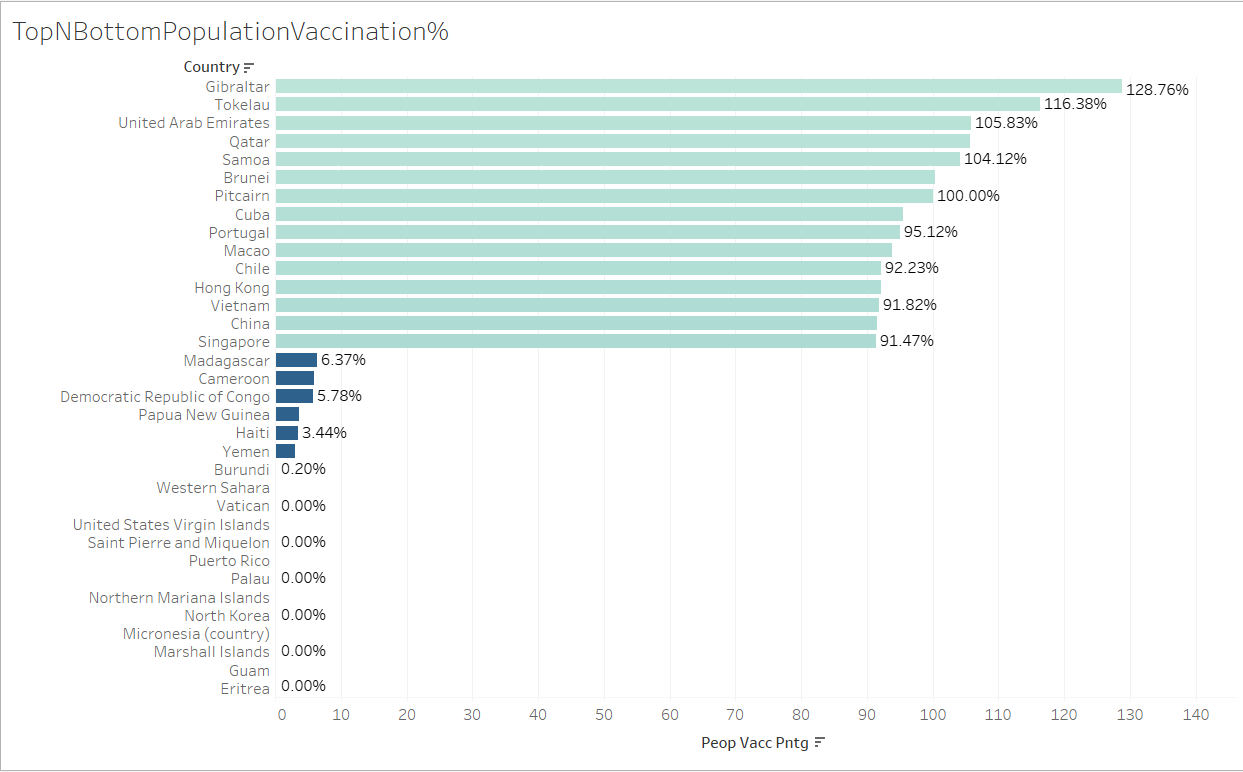
from view\_covid\_data

group by country

order by 2 desc



*Key findings:* When the percentage of people vaccinated calculated over total population of the countries, it is found that countries like Burundi, Yemen, Haiti, Papua New Guinea have lowest i.e., below 0.5%. Whereas Gibraltar, Tokelau, United Arab Emirates, Qatar, Samoa have above 100%.



**14. when did countries started vaccination drive.**

with temp\_tb as

(select RANK() OVER(PARTITION BY country ORDER BY country, rec\_date) as rn,

country, rec\_date as VaccinStartDate

, new\_vacc as vaccinCnt

from view\_covid\_data

where new\_vacc is not null

)

select

ROW\_NUMBER() over(order by VaccinStartDate) as rank\_no,

country,

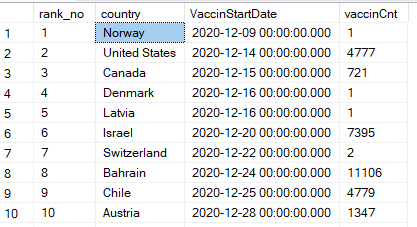
VaccinStartDate,

vaccinCnt

from temp\_tb

where rn = 1

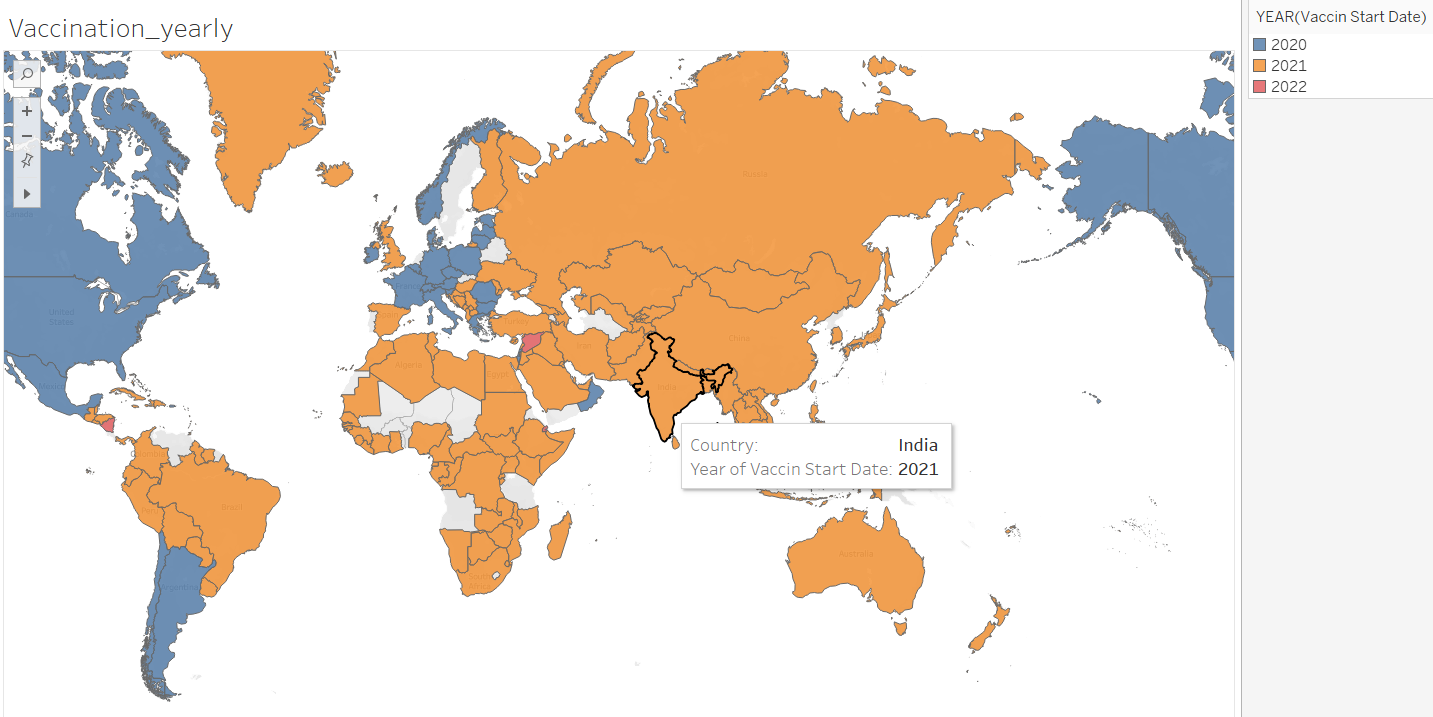
order by VaccinStartDate



*Key findings:* The resultant data shows that Norway and United states have started early than other countries.

Below map shows the top and bottom 5 countries which have started vaccinations first. 

Vaccinations started across globe yearly wise.



**15. Countries which are not vaccinated:**

select country, MAX(cntry\_pop) as TotalPop

,MAX(CAST(tot\_cases as bigint)) as TotalCases

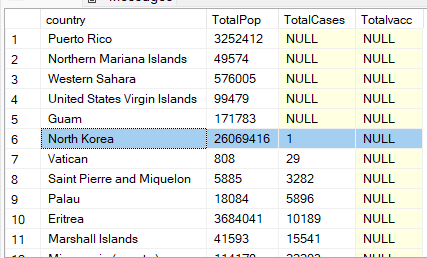
,MAX(CAST(tot\_vacc as bigint)) as Totalvacc

from view\_covid\_data

group by country

having MAX(CAST(tot\_vacc as bigint)) is null

order by 3



*Key findings:* Mostly countries without cases and vaccinations data could be territory of another countries, so their cases and vaccinations count might be included with the dependent countries. We need more data to analyze those locations. Further there are countries with cases and without vaccinations data are like North Korea, Vatican, Saint Pierre and Miquelon, Palau, Eritrea, Marshall Islands, Micronesia (country), which indicates no vaccinations were incorporated.

Now let’s do analysis by gauging the correlation between different attributes like population density by case percentage, diabetes prevalence by deaths.

**16. Countries infection rate by population density:**

Population density refers to measurement of population per unit land area. Let’s understand the impact of population density over covid infection rate.

select country, MAX(cntry\_pop) as Popln

, MAX(pop\_den) as popuDen

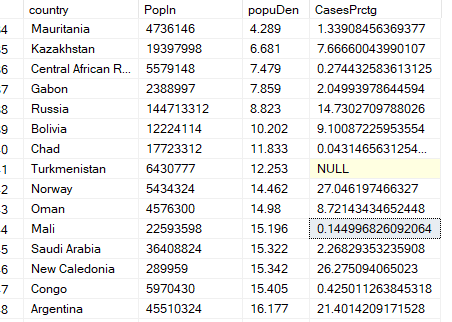
, (MAX(tot\_cases)/MAX(cntry\_pop))\*100 as CasesPrctg

from view\_covid\_data

--and cd.location = 'India'

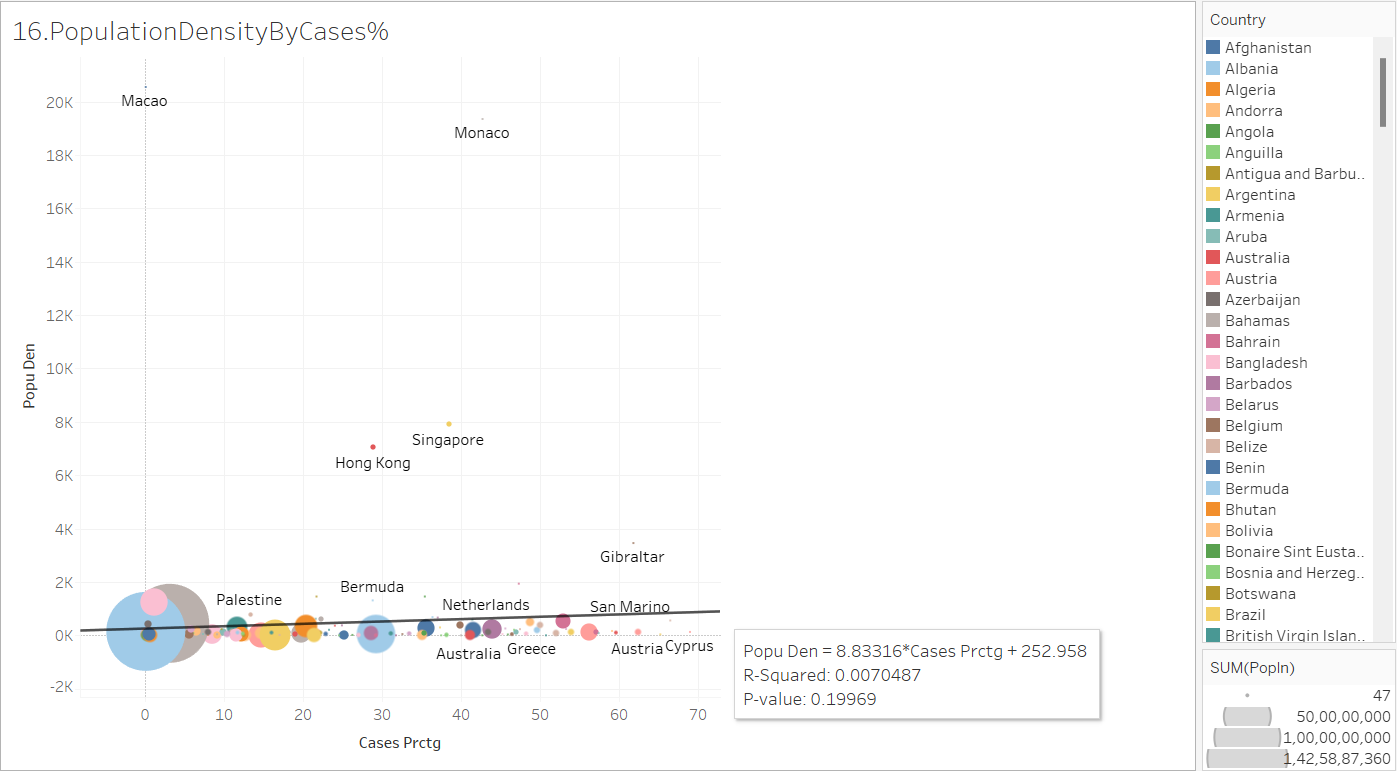
group by country

order by 3

****

*Key findings:* From the data outcome it is observed that cases in the countries is not varying according to population density. More insights on this using scatter plot during visualization can be made.

Studying the impact of population density on the infection rate. Here the colour legend indicates countries where in the size of the bubbles the population of the countries. In the plot, the line is parallel to the x-axis which indicates that there is slight correlation between population density and cases percentage.



**17. Impact of diabetes prevalence over total deaths in countries:**

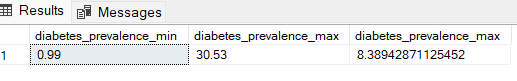
Diabetes prevalence refers to the percentage of people ages 20-79 who have type 1 or type 2 diabetes. It is calculated by adjusting to a standard population age-structure.

select min(diab\_prevalence) diabetes\_prevalence\_min

, max(diab\_prevalence) diabetes\_prevalence\_max

, AVG(diab\_prevalence) diabetes\_prevalence\_max

from view\_covid\_data



It is observed that the range of diabetes prevalence over all countries lies between 0.99% to 30.53%.

Looking at the broader picture of this attribute impact across the globe death rate.

select country, MAX(cntry\_pop) as cntry\_pop

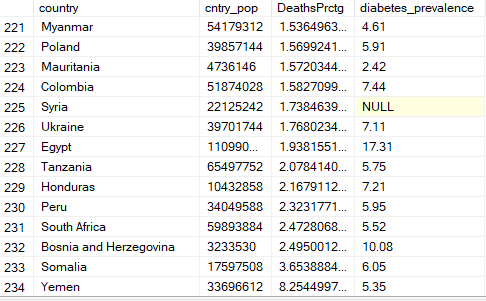
,(MAX(tot\_deaths)/MAX(tot\_cases))\*100 as DeathsPrctg

,MAX(diab\_prevalence) diabetes\_prevalence

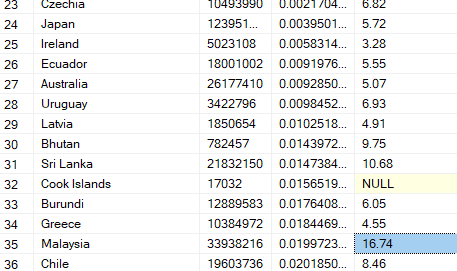
from view\_covid\_data

group by country

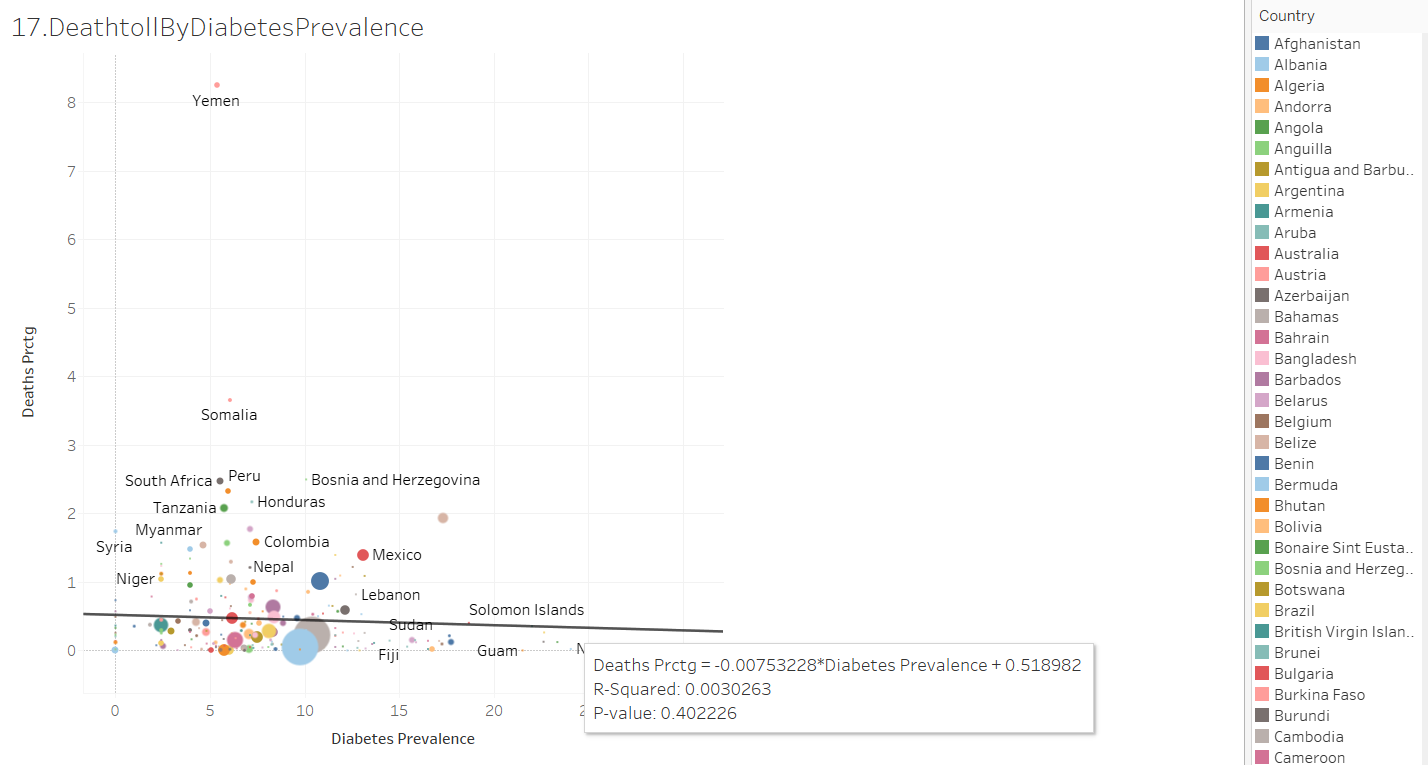
order by 3



The average death rate is around 3.3. But Somalia country have with exceptional high death percentage in spite having below average diabetes prevalence rate.



When we consider the other end values like Japan, Ireland countries have low death rate and below average diabetes prevalence. So, understanding merely by looking at the data is not supporting to the correlations between two attributes. During visualisation we can understand the key points better.



**Visualization:**

Saving the results of all the scenarios into .xlsx files. These files are then used as source file for our visualization in Tableau.

https://public.tableau.com/app/profile/sowjanya.kanike

