COVID-19 DATA ANALYSIS

Introduction

- Though it has subsided by a large extent as of 2024, the COVID-19 pandemic has reshaped the world in unprecedented ways. Our 'COVID-19 Data Analysis' project delves into a comprehensive exploration of the pandemic's effects, focusing on data-driven insights to understand its impact on different nations and how they have reacted to it.

As the global community continues to navigate the complexities of the pandemic, my project embarks on a data-driven journey to uncover the multifaceted aspects of the COVID-19 crisis. Through meticulous data analysis and visualization, Iseek to uncover patterns, trends, and lessons that can guide decision-making.

My mission is to harness the power of data to go beyond headlines and delve into the nuances of the pandemic's effects across countries. By examining key metrics, Iaim to provide a comprehensive overview of how nations have grappled with this global challenge.

Gathering Data

Igathered and downloaded the Covid-19 dataset from the site 'Our world in data'

(link : https://ourworldindata.org/covid-deaths).

- The original dataset contains over a 90 columns and several thousand rows of data,
- containing data of many countries in columns like 'Total cases of infection', 'change in vaccination percentage', etc. to name a few. This data is spread across several decades for each country, hence helping us to analyze each country's covid performance with more clarity

Data Cleaning

The data was stored in a .csv format. Although the data was arranged quite beautifully, in many places there were many blank spaces. These blank spaces meant that there was no particular data available for that row and the specific column.

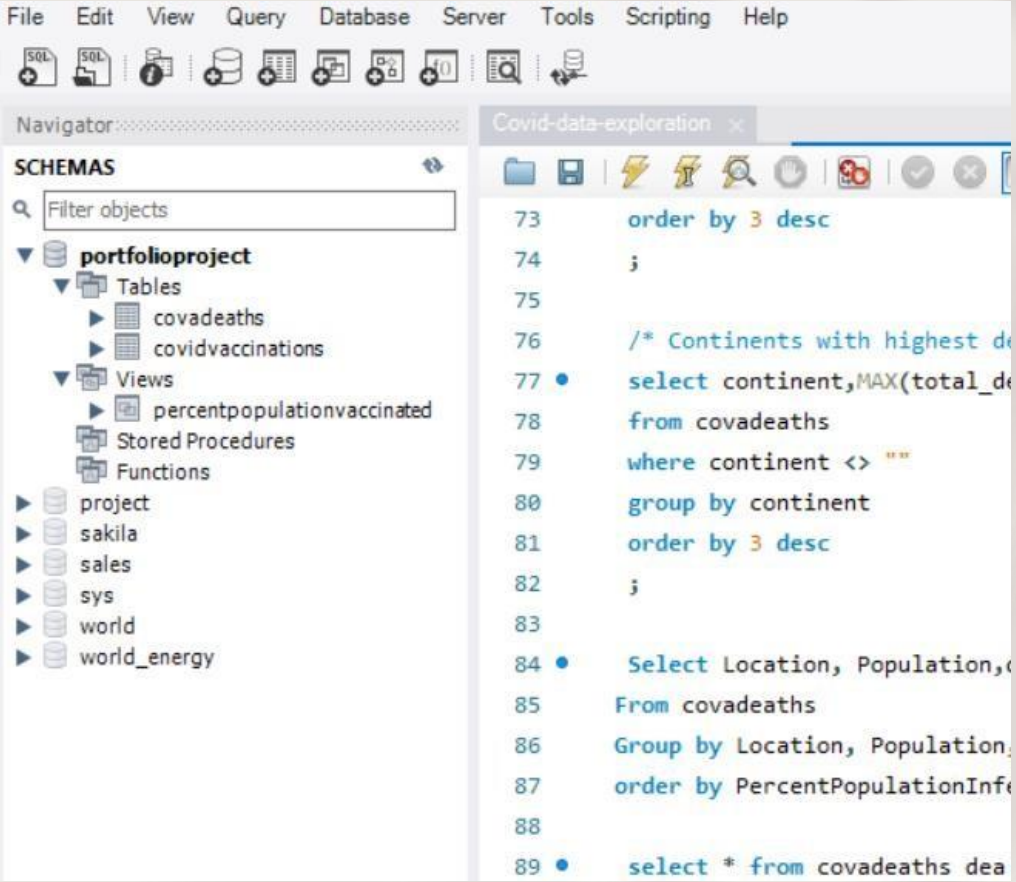
Now the problem is, since we have to connect this .csv file to our MySQL server, we cannot afford to have blank spaces, or , empty cells. Because in this case, it will read them just as a series of comas and MySQL may not understand that these are actually empty cells.

But before that, I split the big table into two halves, one table for the Covid Deaths, and the other for data related to Covid Vaccinations.

Hence to prevent any errors, I first filled all the blank cells with zeroes. After that, I used the 'load infile' command to upload both the .csv files into the MySQL server and then , we had two tables with all the data contained within them.

We had split the original table into two parts, one for storing data exclusively related to covid deaths and other for covid vaccinations.

As you can see, we have two tables here, one is 'covadeaths' and the other is 'covidvaccinations'.



Writing Queries

- So after successfully loading our .csv files into MySQL , it's time to gain some insights on our data and answer some questions related to energy distribution and energy usage of nations.

- We answer some questions and looked into things like :

1) What are the total no. of infections , to number of deaths globally? What is the global death percentage?
2) What is the trend of percent population affected per country over time?
3) What is the continent wise count of total deaths related to Covid? Which continent has suffered the most?
4) What is the percentage of population infected by the virus, per country?
5) Are the covid deaths anyhow related or influenced by the cardiovascular deaths?

- To answer these questions we first wrote relevant queries in MySQL and extracted the information. After extracting the relevant information, we put them onto excel files which were then in turn connected to Tableau, a BI tool which helps us to visualize and present our data in a more beautiful way so that we can explain it better to relevant stakeholders.

Some queries :

Navigator

SCHEMAS

Filter objects

- portfolioproject
  - Tables
    - covadeaths
    - covidvaccinations
  - Views
    - percentpopulationvaccinated
  - Stored Procedures
  - Functions
- project
- sakila
- sales
- sys
- world
- world_energy

Administration Schemas

Information

No object selected

Covid-data-exploration

Limit to 50000 rows

```
 order by 1,2
 ;

 /*Total Cases, death rates and deaths acrosss the world*/

Select SUM(new_cases) as total_cases, SUM(new_deaths) as total_deaths, SUM(new_deaths)/SUM(New_Cases)*100 as DeathPercentage
From covadeaths
where continent is not null
order by 1,2;

/* Countries with highest infecton rate compared to population --->*/

select location,population,MAX(total_cases),max(total_cases/population)*100 as Total_infected_Percentage_of_population
 from covadeaths
 where continent is not null
 group by location,population
 order by 4 desc
 ;

 /* Countries with highest death rate compared to population --->*/
 select location ,MAX(total_deaths),max(total_deaths/population)*100 as Total_death_Percentage_of_population
 from covadeaths
 where continent <> ""
 group by location
 order by 3 desc
 ;
```

```
order by 1,2
;

/*Total Cases, death rates and deaths acrosss the world*/

Select SUM(new_cases) as total_cases, SUM(new_deaths) as total_deaths, SUM(new_deaths)/SUM(New_Cases)*100 as DeathPercentage
From covadeaths
where continent is not null
order by 1,2;

/* Countries with highest infecton rate compared to population --->*/

select location,population,MAX(total_cases),max(total_cases/population)*100 as Total_infected_Percentage_of_population
 from covadeaths
 where continent is not null
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_cases	total_deaths	DeathPercentage
3253410813	29058502	0.8932

Covid-data-exploration

Limit to 50000 rows

```
From covadeaths
where continent is not null
order by 1,2;

/* Countries with highest infecton rate compared to population --->*/

select location,population,MAX(total_cases),max(total_cases/population)*100 as Total_infected_Percentage_of_population
 from covadeaths
 where continent is not null
 group by location,population
 order by 4 desc
 ;

 /* Countries with highest death rate compared to population --->*/
 select location ,MAX(total_deaths),max(total_deaths/population)*100 as Total_death_Percentage_of_population
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

location	population	MAX(total_cases)	Total_infected_Percentage_of_population
Cyprus	896007	660854	73.7600
San Marino	33690	24326	72.2100
Brunei	449002	308777	68.7700
Austria	8939617	6081287	68.0300
Faeroe Islands	53117	34658	65.2500
Slovenia	2119843	1344559	63.4300
Gibraltar	32677	20550	62.8900
Martinique	367512	230354	62.6800
South Korea	51815808	32256154	62.2500
Andorra	79843	48015	60.1400
Jersey	110796	66391	59.9200
Luxembourg	647601	382785	59.1100

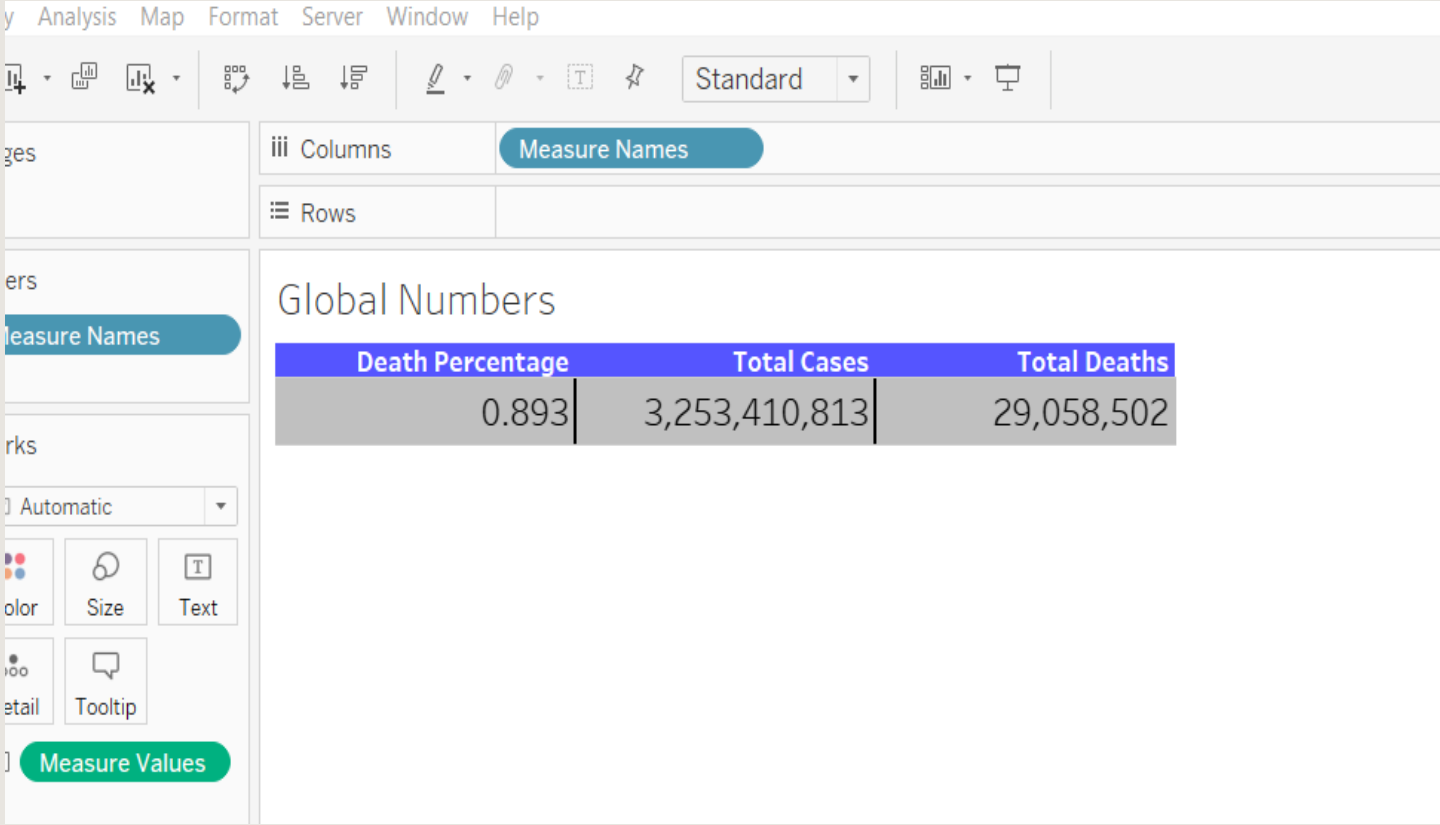
Result 2

Connecting data source to Tableau

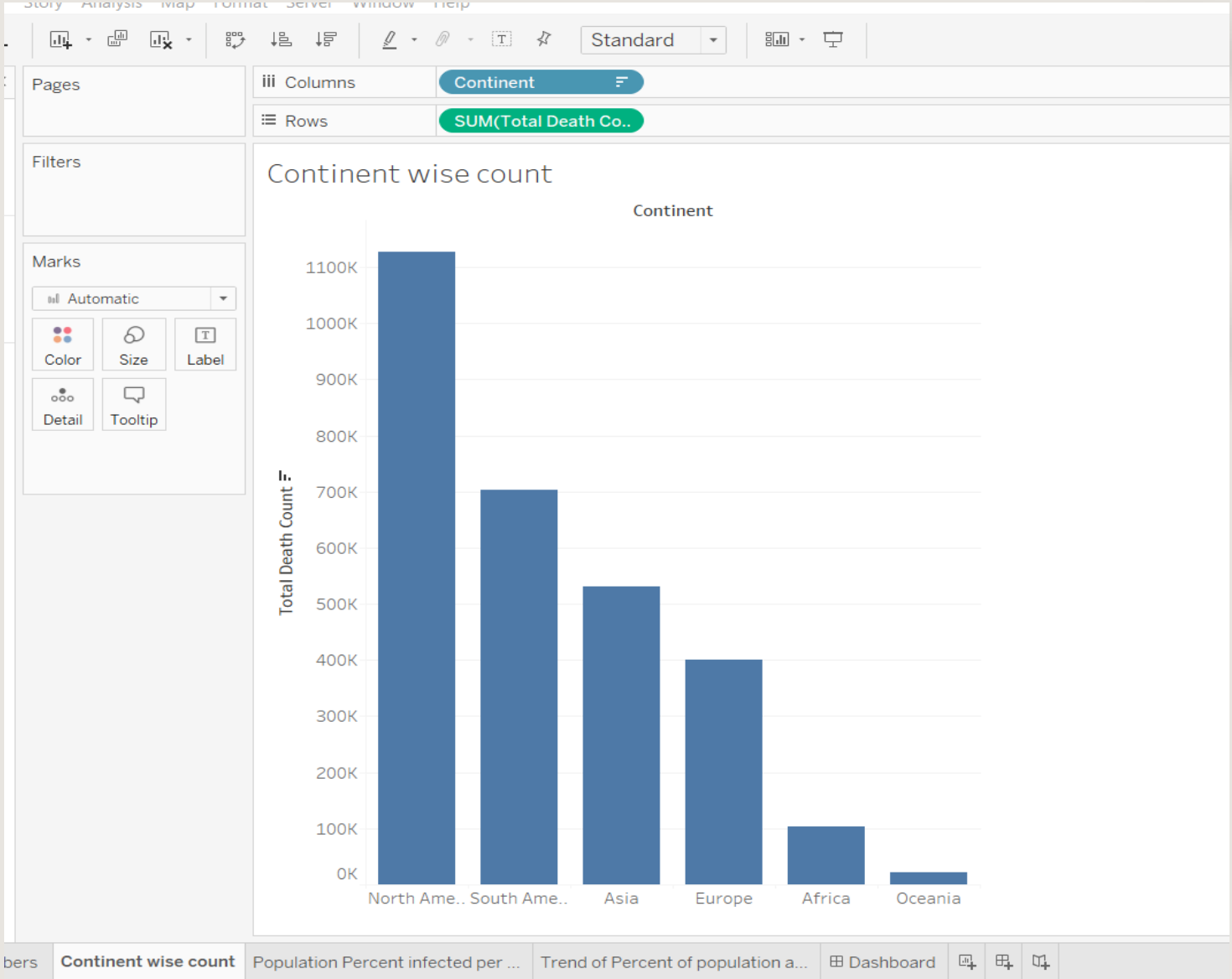
- After getting all the desried results and outputs from MySQL, we copy the output from MySql and paste them on excel files. This is done because Tableau public does not support importing data directly from sql files.

Tableau visualizations :

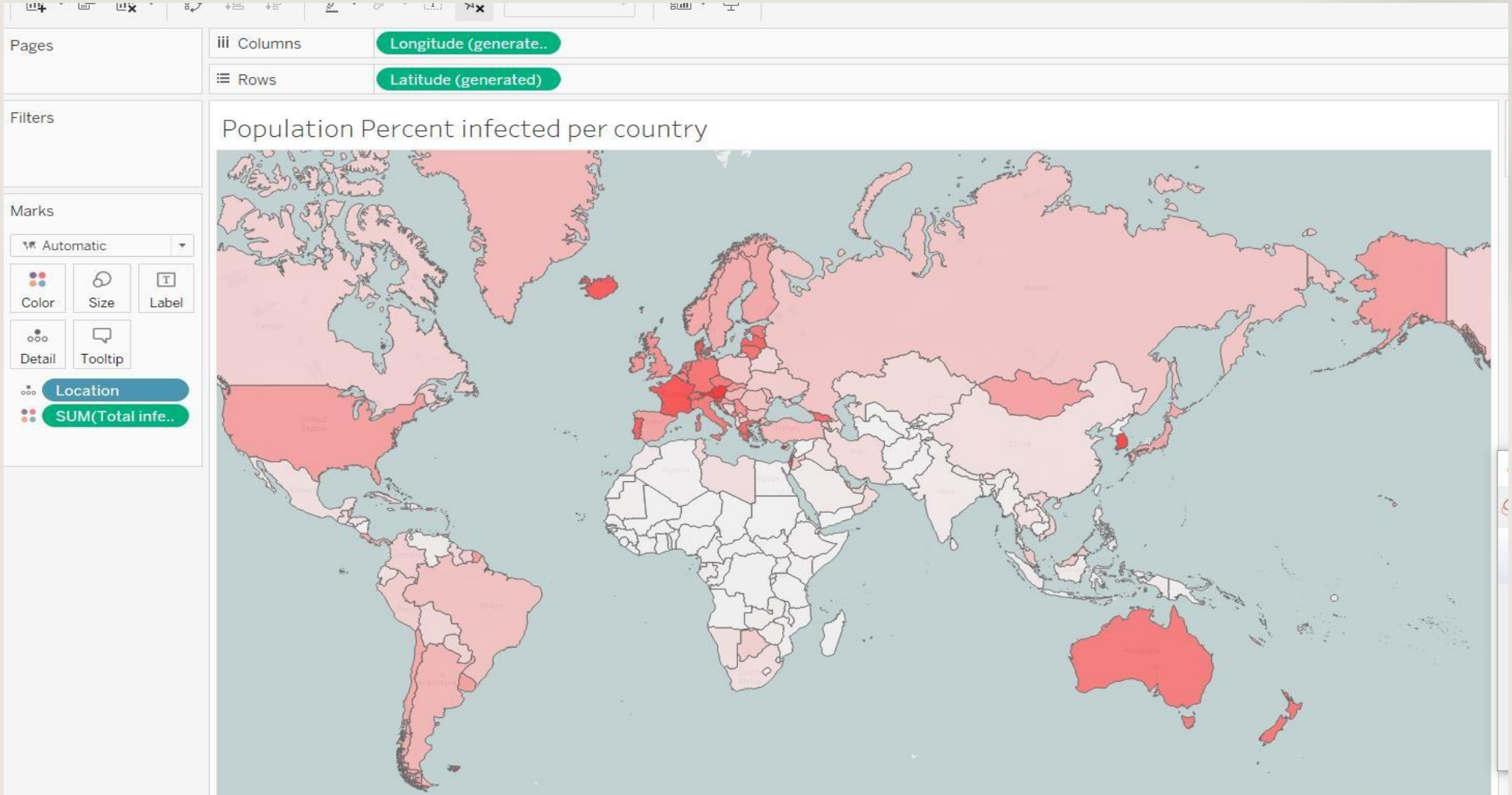
What are the total no. of infections , to number of deaths globally? What is the global death percentage?



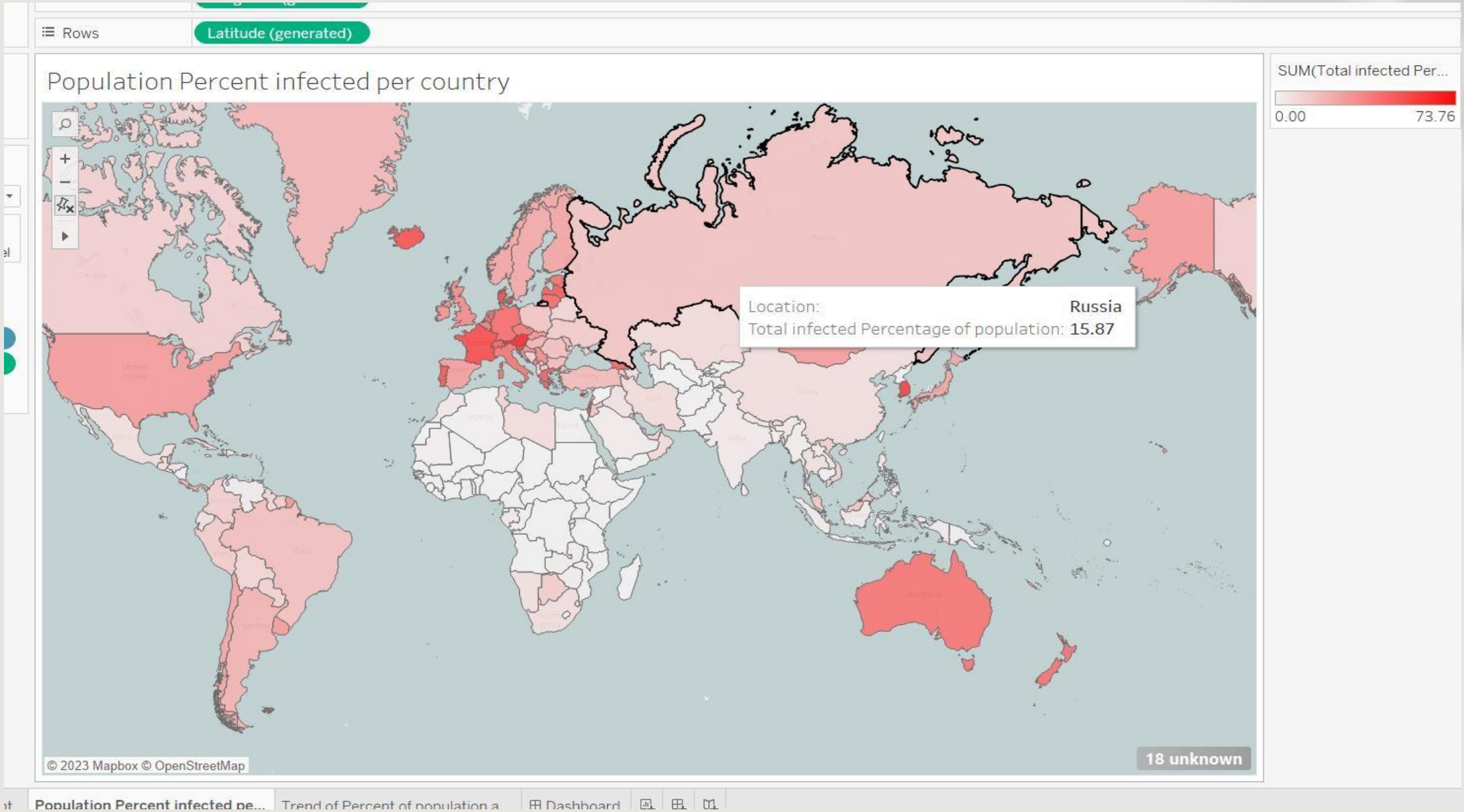
- What is the continent wise count of total deaths related to Covid? Which continent has suffered the most?



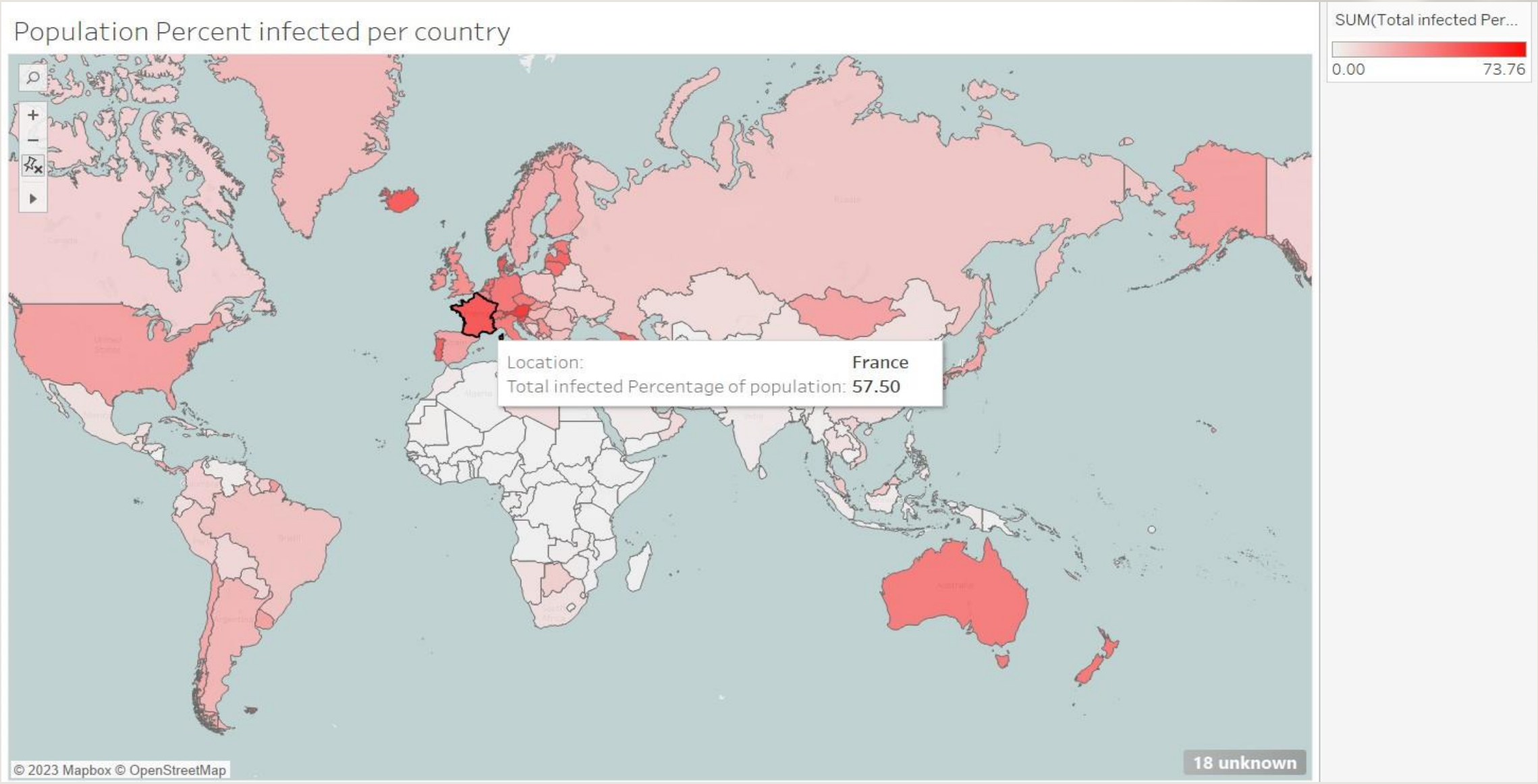
- **What is the percentage of population infected by the virus, per country?**
- For this viz, we have plotted the entire world map and one can view the percentage infected of each country by simply hovering/tapping on that particular country.



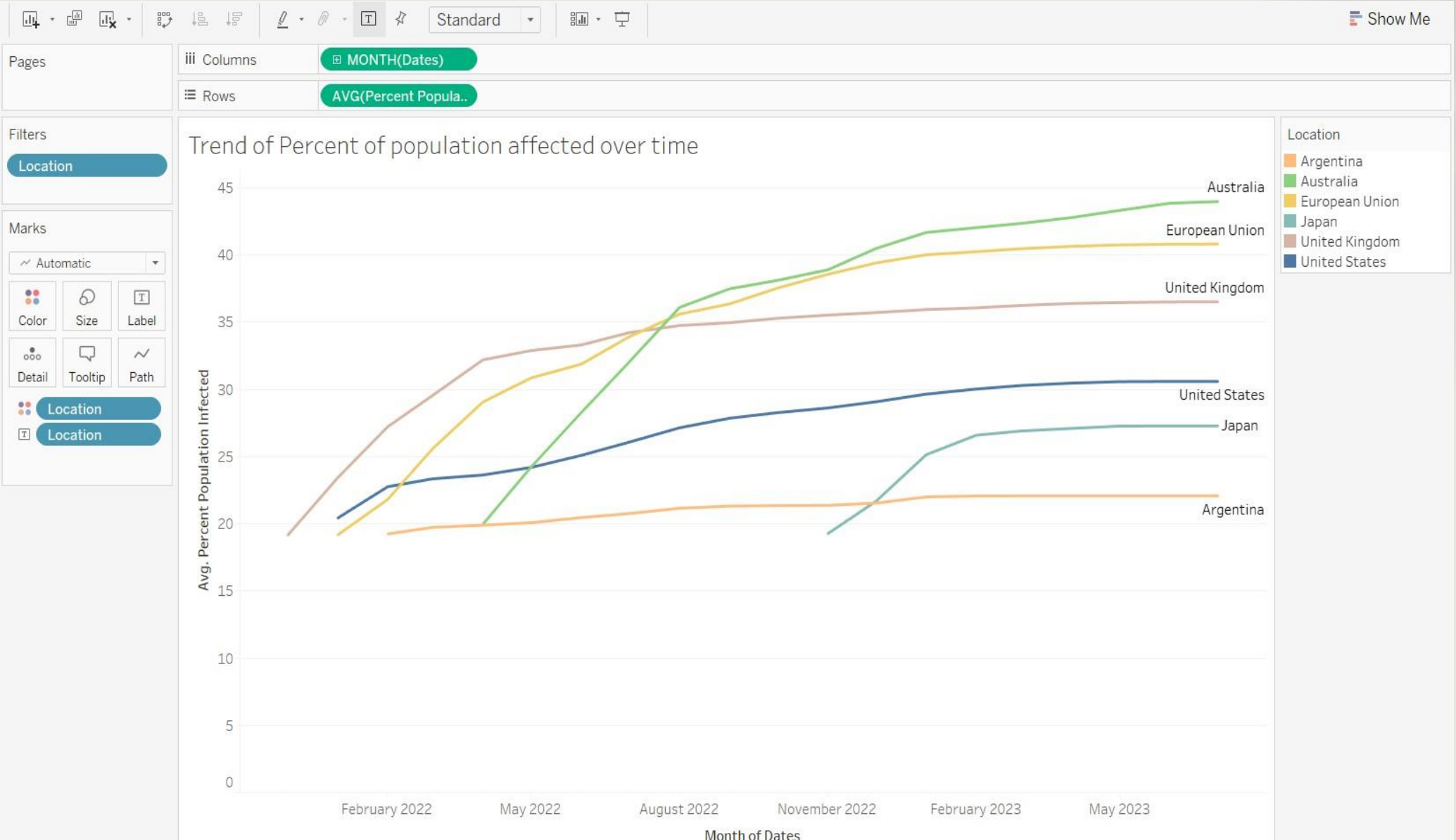
- E.g., for Russia



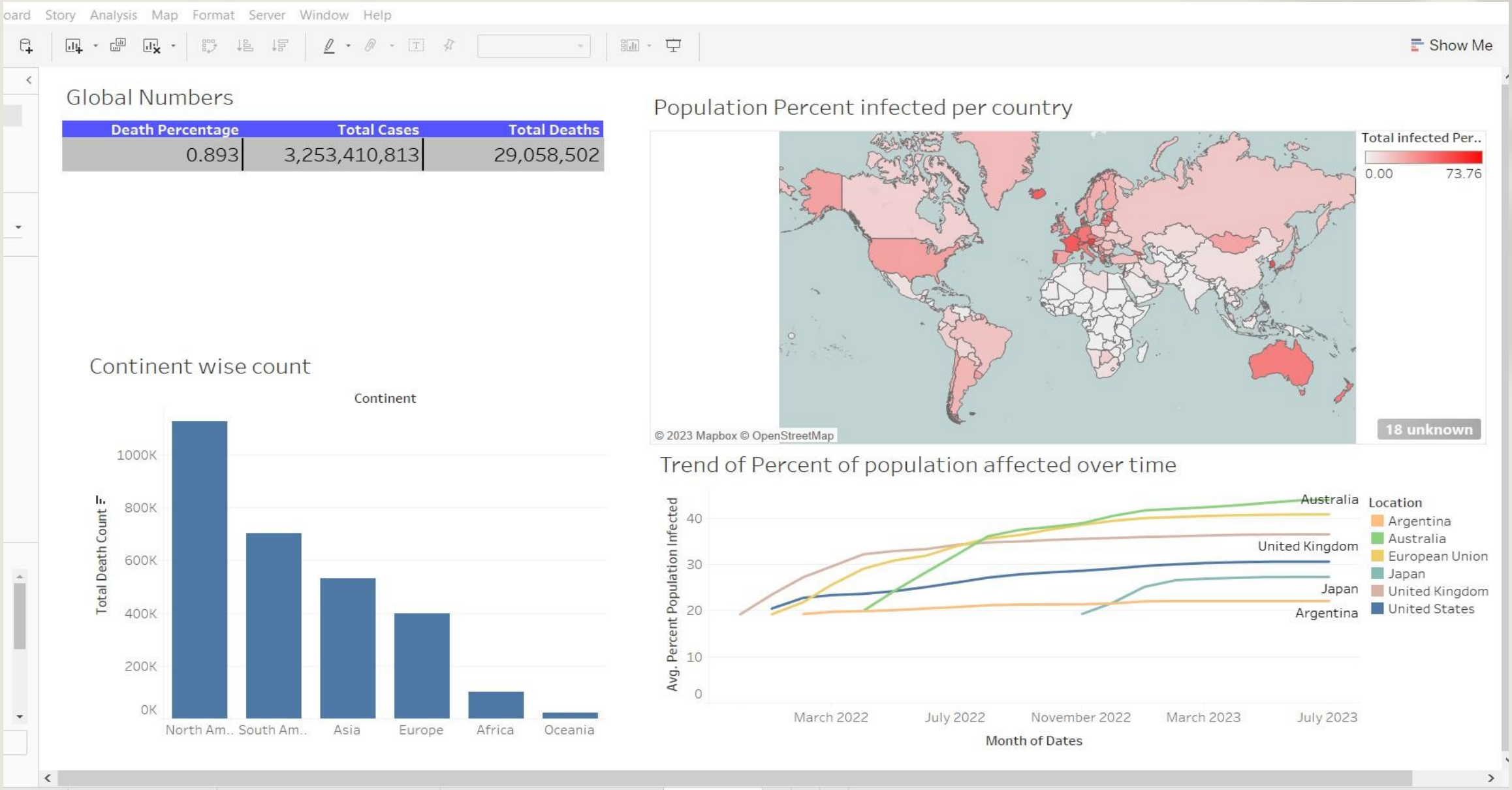
- For France,


Population Percent infected per country
SUM(Total infected Per...
0.00
73.76
Location: France
Total infected Percentage of population: 57.50
© 2023 Mapbox © OpenStreetMap
18 unknown

- What is the trend of percent of population infected over time per country, over the past one year?



- And finally, presenting all our viz. Together in the **Dashboard** :



Are the covid deaths influenced by the cardiovascular diseases/deaths?

1) To answer this question, we take a rather more mathematical approach. We perform Simple linear regression using Excel, and find out.
2) First, we took out only the relevant data from the original huge table. Mainly, we needed two columns, the ‘total_covid_deaths_per_million’ column and the ‘cardiovascular death rate’ column.
3) The ‘cardiovascular death rate’ column had figures for deaths per lakh. Hence, a new column ‘cdsv_deathreate_permillion’ was created for having the death figures in millions for a better comparison and calculations.
4) This is how the original , unfiltered data looks :

C5099 | Algeria

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	iso_code	continent	locatio	date	total_cases	new_cases	ew_cases_smoothe	total_deaths	ew_death	eaths_smo	ases_per_	ases_per_	w_cases_smoothed_per_mill	total_deaths_
2	AFG	Asia	Afghanista	2020-01-05		0			0			0		
3	AFG	Asia	Afghanista	2020-01-06		0			0			0		
4	AFG	Asia	Afghanista	2020-01-07		0			0			0		
5	AFG	Asia	Afghanista	2020-01-08		0			0			0		
6	AFG	Asia	Afghanista	2020-01-09		0			0			0		
7	AFG	Asia	Afghanista	2020-01-10		0	0		0	0		0	0	
8	AFG	Asia	Afghanista	2020-01-11		0	0		0	0		0	0	
9	AFG	Asia	Afghanista	2020-01-12		0	0		0	0		0	0	
10	AFG	Asia	Afghanista	2020-01-13		0	0		0	0		0	0	
11	AFG	Asia	Afghanista	2020-01-14		0	0		0	0		0	0	
12	AFG	Asia	Afghanista	2020-01-15		0	0		0	0		0	0	
13	AFG	Asia	Afghanista	2020-01-16		0	0		0	0		0	0	
14	AFG	Asia	Afghanista	2020-01-17		0	0		0	0		0	0	
15	AFG	Asia	Afghanista	2020-01-18		0	0		0	0		0	0	
16	AFG	Asia	Afghanista	2020-01-19		0	0		0	0		0	0	
17	AFG	Asia	Afghanista	2020-01-20		0	0		0	0		0	0	
18	AFG	Asia	Afghanista	2020-01-21		0	0		0	0		0	0	
19	AFG	Asia	Afghanista	2020-01-22		0	0		0	0		0	0	
20	AFG	Asia	Afghanista	2020-01-23		0	0		0	0		0	0	
21	AFG	Asia	Afghanista	2020-01-24		0	0		0	0		0	0	
22	AFG	Asia	Afghanista	2020-01-25		0	0		0	0		0	0	
23	AFG	Asia	Afghanista	2020-01-26		0	0		0	0		0	0	
24	AFG	Asia	Afghanista	2020-01-27		0	0		0	0		0	0	
25	AFG	Asia	Afghanista	2020-01-28		0	0		0	0		0	0	
26	AFG	Asia	Afghanista	2020-01-29		0	0		0	0		0	0	
27	AFG	Asia	Afghanista	2020-01-30		0	0		0	0		0	0	

Sheet1 Sheet2

5) After that we selected and filtered only the columns and data that we need for our particular analysis. We added an extra column ‘Header’ to only keep the last data available date data for each country.This is how the data looks now :

J22374

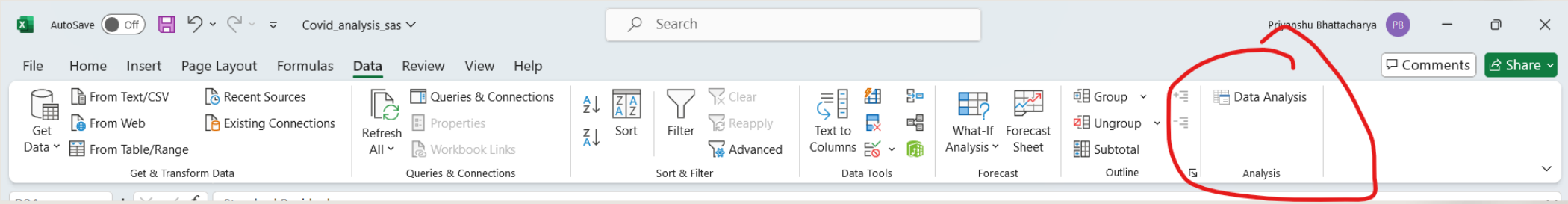
	A	B	C	D	E	F	G	H	I
1	contine	location	date	total_d	gdp_pe	cardiov	cdsv_de	diabete	Header
1598	Asia	Afghanista	2024-05-1	194.462	1803.987	597.029	5970.29	9.59	Last
4792	Europe	Albania	2024-05-1	1268.331	11803.43	304.195	3041.95	10.08	Last
6389	Africa	Algeria	2024-05-1	153.241	13913.84	278.364	2783.64	6.73	Last
7986	Oceania	American S	2024-05-1	767.581		283.75	2837.5		Last
9583	Europe	Andorra	2024-05-1	1991.408		109.135	1091.35	7.97	Last
1180	Africa	Angola	2024-05-1	54.427	5819.495	276.045	2760.45	3.94	Last
2777	North Ame	Anguilla	2024-05-1	755.81	0	0	0		Last
4374	North Ame	Antigua an	2024-05-1	1556.968	21490.94	191.511	1915.11	13.17	Last
5975	South Ame	Argentina	2024-05-1	2875.392	18933.91	191.032	1910.32	5.5	Last
7572	Asia	Armenia	2024-05-1	3156.658	8787.58	341.01	3410.1	7.11	Last
9169	North Ame	Aruba	2024-05-1	2742.84	35973.78	0	0	11.62	Last
2374	Oceania	Australia	2024-05-1	964.037	44648.71	107.791	1077.91	5.07	Last
3971	Europe	Austria	2024-05-1	2520.69	45436.69	145.183	1451.83	6.35	Last
5568	Asia	Azerbaijan	2024-05-1	999.51	15847.42	559.812	5598.12	7.11	Last
7165	North Ame	Bahamas	2024-05-1	2068.348	27717.85	235.954	2359.54	13.17	Last
8762	Asia	Bahrain	2024-05-1	1043.31	43290.71	151.689	1516.89	16.52	Last
0367	Asia	Bangladesl	2024-05-2	0	3523.984	298.003	2980.03	8.38	Last
1964	North Ame	Barbados	2024-05-1	2105.48	16978.07	170.05	1700.5	13.57	Last
3561	Europe	Belarus	2024-05-1	746.516	17167.97	443.129	4431.29	5.18	Last
5158	Europe	Belgium	2024-05-1	2946.056	42658.58	114.898	1148.98	4.29	Last
6755	North Ame	Belize	2024-05-1	1697.571	7824.362	176.957	1769.57	17.11	Last
8352	Africa	Benin	2024-05-1	12.207	2064.236	235.848	2358.48	0.99	Last
9949	North Ame	Bermuda	2024-05-1	2569.813	50669.32	139.547	1395.47	13	Last
1546	Asia	Bhutan	2024-05-1	26.839	8708.597	217.066	2170.66	9.75	Last
3143	South Ame	Bolivia	2024-05-1	1831.38	6885.829	204.299	2042.99	6.89	Last
4740	North Ame	Bonaire Si	2024-05-1	1515.6	0	0	0		Last

Sheet4 | Sheet1 | Sheet6 | Sheet5 | last column data | Sheet7 | Regression data | First | Sheet3

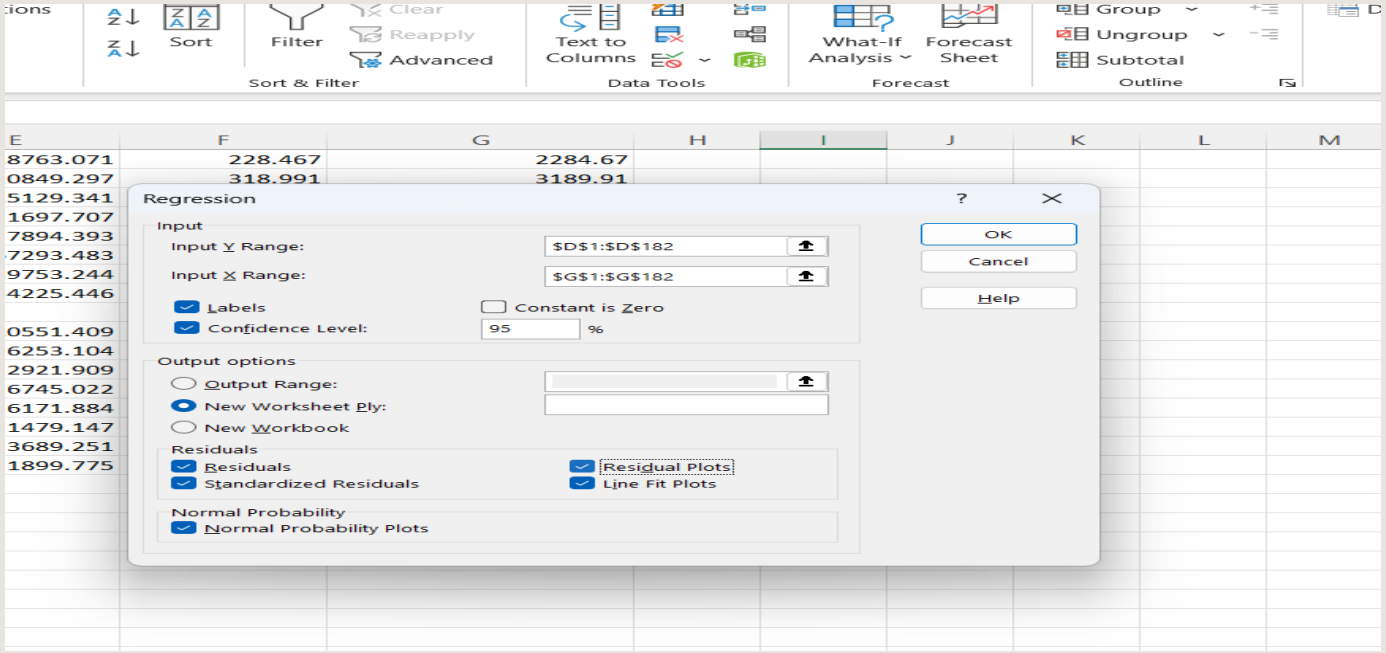
Ready | Filter Mode | Accessibility: Investigate | 100%

8) Now that we have our linear equation of y = -0.19826x + 2892, we will do a proper simple linear regression to verify the same as well get more info on our data which will give us a clear understanding to answer our business problem.

9) We go to the 'Data' Tab in the excel ribbon above and select 'Regression'.



10) After going on 'Regression', we input our x and y column values, and also select the data that we want to see. In our case, this is how we do the entries :



11) This generates our simple linear regression data and three plots representing the various relationships between the variables in our data.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.215175
R Square	0.0463
Adjusted R	0.040972
Standard E	1289.367
Observatic	181

ANOVA

	df	SS	MS	F	ignificance F
Regressior	1	14446964	14446964	8.690079	0.003626
Residual	179	2.98E+08	1662466		
Total	180	3.12E+08			

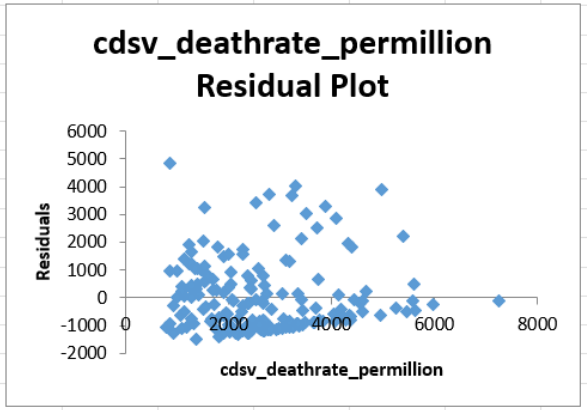
	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	1829.504	230.9926	7.920187	2.42E-13	1373.685	2285.323	1373.685	2285.323
cdsv_deat	-0.2336	0.079244	-2.94789	0.003626	-0.38998	-0.07723	-0.38998	-0.07723

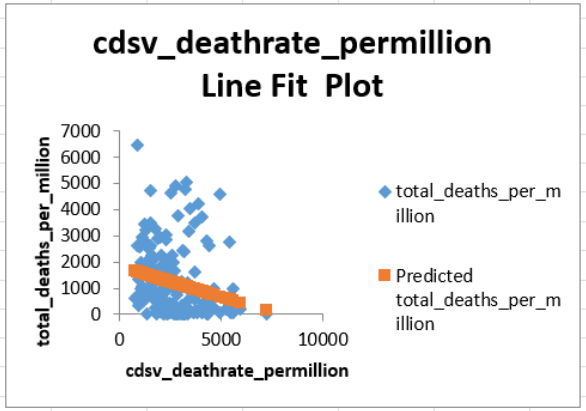
RESIDUAL OUTPUT

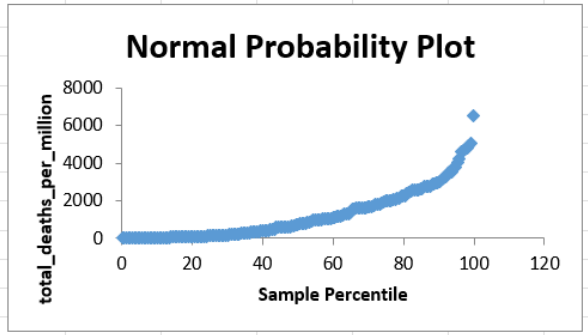
Observatior	tal_deaths_	Residuals	dard Residuals
1	434.8249	-240.363	-0.18694
2	1118.895	149.4363	0.116222
3	1179.237	-1026	-0.79796

PROBABILITY OUTPUT

Percentile	eaths_per_million
0.276243	1.164
0.828729	10.946
1.381215	12.019







Observations and Inferences

1) Multiple R : We have a low multiple R. It basically means, that the correlation between our x and y variables is not very high.

2) R square : We have a R square of 0.0463 or 4.63% only, which means that only 4.63% of the variance in y can be explained by x.

3) Significance F : As we had taken 95% confidence interval, this value must be less than 0.05 to be considered significant. In our case, it is around 0.0036, hence it can be considered significant.

4) Coefficients : We have the intercept coefficient as **1829.5 (b)** and the slope coefficient as **-0.2336(m)** for our linear equation .

5) Hence, we can say that since our slope is negative, according to this data, with a unit increase in X, our Y decreases by 0.2336 units. Hence, total deaths is not increasing due to increase with cardiovascular deaths.

6) Looking at the plots, if we look at the residual plot, the upper half and lower half of the graph are not evenly distributed, hence, the model is not as ideal as we would want it to be.

7) Looking at the line fit plot, again, we saw the orange line or the 'predicted total deaths' passing through the actual data points in a downward direction. This again weakens the point that these two variables are directly related.

8) Finally, if we look at the normal probability graph, we see a curved line spread across the percentile. Ideally , it should be a straight upwards line as it indicates that the predicted Y values and the actual Y values , both are increasing together. Hence, this graph also negates a positive relation between the two.

Final Conclusion : Since our R square values are low (4.63%) , the slope coefficient is negative (-0.2336), and all the three graphs also do not match or come near the ideal scenario, we can safely conclude that based on this data, the total deaths caused by covid is not influenced by the cardiovascular deaths across all countries.

Conclusion

- Hence, we saw how using Excel,SQL and Tableau can be extremely helpful
  in cleaning data,organisingdata, retreiving specifi c information and finally presenting them in a visually appealing manner to the stakeholders.
- We got insights of how covid has impacted different countries around the world .