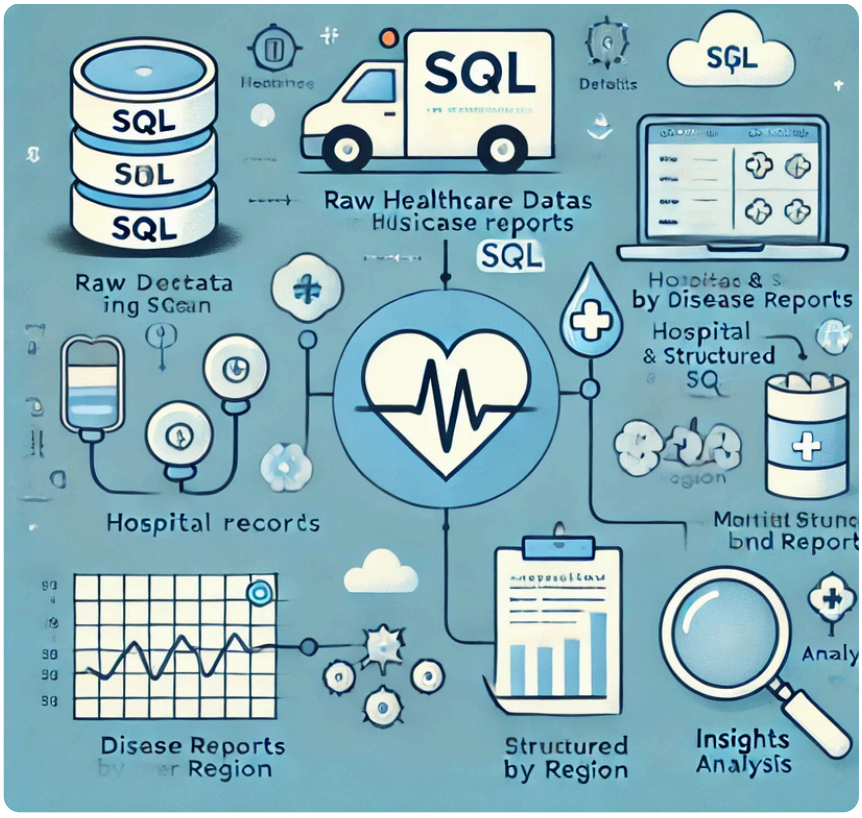
Healthcare Analytics:
Insights into Global Causes of Death

Project by: Prachit Gopidwad

Date: Jan 29, 2025

This project leverages SQL to analyze global mortality data, identifying patterns and trends across diseases, regions, and time periods. By cleaning, transforming, and analyzing data using PostgreSQL, we aim to uncover key insights that support public health initiatives and resource allocation. The findings will help in identifying high-risk diseases and regions needing intervention. Additionally, this project highlights the importance of data-driven healthcare solutions in optimizing medical resources. Through SQL-based analytics, we can efficiently assess healthcare trends and support evidence-based decision-making for better public health outcomes.



Background of study

Background:

- Understanding these causes is essential for policymakers, healthcare professionals, and researchers to address health disparities and improve outcomes.

Goals

- To provide a comprehensive understanding of global mortality trends.
- To support evidence-based decision-making for health interventions.

Objectives

- To analyze global causes of death across countries and over time using a comprehensive dataset.
- To identify patterns, trends, and areas of concern that can provide actionable insights to inform healthcare policies.

Scope:

- Data Source: A dataset containing annual mortality statistics by cause, country, and year.
- Tools Used: PostgreSQL for data cleaning, transformation, and querying.

Importance of the Project:

- Public Health Impact: Helps identify priority areas for healthcare interventions.
- Resource Allocation: Informs efficient allocation of healthcare resources.
- Data-Driven Decisions: Empowers stakeholders to make informed decisions based on evidence.

Data Exploration & Preparation

Data Overview: 🌐 **cause_of_deaths.csv**

- The dataset contains **6120** records.
- The dataset contains the following columns:

- country_territory
- code
- year
- meningitis
- alzheimers_disease_and_related_dementias
- parkinsons_disease
- nutritional_deficiencies
- malaria
- drowning
- interpersonal_violence
- maternal_disorders
- hiv_aids
- drug_use_disorders
- tuberculosis
- cardiovascular_diseases
- lower_respiratory_infections
- neonatal_disorders
- alcohol_use_disorders
- self-harm
- exposure_to_natural_forces
- diarrheal_diseases
- environmental_exposure_to_heat_and_cold
- neoplasm
- conflict_and_terrorism
- diabetes _mellitus
- chronic_kidney_diseases
- poisonings
- protein_energy_malnutrition
- road_injuries
- chronic_respiratory_diseases
- cirrhosis_liver_diseases
- diestive_diseases
- fire_heat_and_hot_substances
- acute_hepatitis

Importing Data into PostgreSQL:

- Creating Table:

```
CREATE TABLE cause_of_deaths (country_territory VARCHAR(30),code VARCHAR(10),
year INT,meningitis INT,alzheimers_dementia INT,parkinsons INT,nutritional_deficiencies INT,
malaria INT,drowning INT,interpersonal_violence INT,maternal_disorders INT,
hiv_aids INT,drug_use_disorders INT,tuberculosis INT,cardiovascular_diseases INT,
lower_respiratory_infections INT,neonatal_disorders INT,alcohol_use_disorders INT,
self_harm INT,exposure_to_nature INT,diarrheal_diseases INT,
environmental_exposure INT,neoplasms INT,conflict_terrorism INT,diabetes_mellitus INT,
chronic_kidney_disease INT,poisonings INT,protein_energy_malnutrition INT,road_injuries INT,
chronic_respiratory_diseases INT,cirrhosis_liver_diseases INT,
digestive_diseases INT,fire_heat_substances INT,acute_hepatitis INT);
```

- Steps For Importing Data Into PostgreSQL:
  - Right click on table name cause_of_deaths
  - Select on import data
  - Select .csv file path
  - Select Import

Challenges Identified:

- **Missing Values:** Some rows contain null values for specific diseases.
- **Duplicates:** Identified potential duplicates in country-year entries.

SQL Cleaning Process:

- Replacing missing values with 0 for numerical columns:

```
UPDATE cause_of_deaths
SET meningitis = 0
WHERE meningitis IS NULL;
```

- Removing duplicate rows based on **country_territory** and **year**:

```
DELETE FROM cause_of_deaths
WHERE id NOT IN (SELECT MIN(id)
FROM cause_of_deaths
GROUP BY country_territory, year);
```

Adding total_deaths Column:

```
ALTER TABLE cause_of_deaths
ADD COLUMN total_deaths int;
```

```
Update cause_of_deaths
set total_deaths = meningitis +alzheimers_dementia +parkinsons +nutritional_deficiencies
+malaria+drowning +interpersonal_violence maternal_disorders +hiv_aids + drug_use_disorders +
tuberculosis +cardiovascular_diseases +lower_respiratory_infections +neonatal_disorders +
alcohol_use_disorders +self_harm +exposure_to_nature +diarrheal_diseases +
environmental_exposure +neoplasms +conflict_terrorism +diabetes_mellitus +
chronic_kidney_disease +poisonings +protein_energy_malnutrition +road_injuries +
chronic_respiratory_diseases +cirrhosis_liver_diseases +digestive_diseases
+fire_heat_substances +acute_hepatitis;
```

Creating a VIEW for further help in analysis

- VIEW must include total death count of each disease

```
create view sum_and_union_of_diseases as
SELECT 'meningitis' as disease, SUM(meningitis) as total_deaths
FROM cause_of_deaths
UNION ALL
SELECT 'alzheimers_dementia', SUM(alzheimers_dementia)
FROM cause_of_deaths
UNION ALL
SELECT 'parkinsons', SUM(parkinsons)
FROM cause_of_deaths
UNION ALL
SELECT 'nutritional_deficiencies', SUM(nutritional_deficiencies)
FROM cause_of_deaths
UNION ALL
/*(Repeat with all disease columns)*/
SELECT 'acute_hepatitis', SUM(acute_hepatitis)
FROM cause_of_deaths
```

Outcome:

- A clean and consistent dataset ready for analysis and visualization.
- Issues such as missing values and duplicate entries have been resolved.

Analysis & Insights

Exploratory Analysis:

- How many countries are included in data?

```sql
select count
(distinct code)
as num_of_countries
from cause_of_deaths;
```

Data Output | Messages | Notifications

	num_of_countries bigint
1	204

The dataset Includes 204 countries including developed, developing and Under- developed countries.

Result:- 1_num_of_countries.csv

- Data from which years are included ?

```sql
select
distinct year
from cause_of_deaths
order by year;
```

Data Output | Messages | Notifications

	year integer
1	1990
2	1991
3	1992
4	1993
5	1994

Data from 30 years (from 1990 to 2019) is included in dataset.

Result:- 2_years.csv

Descriptive Analysis:

- What is the total number of deaths for each disease globally?

```
296 SELECT disease, total_deaths
297 FROM sum_and_union_of_diseases
298 ORDER BY total_deaths desc;
```

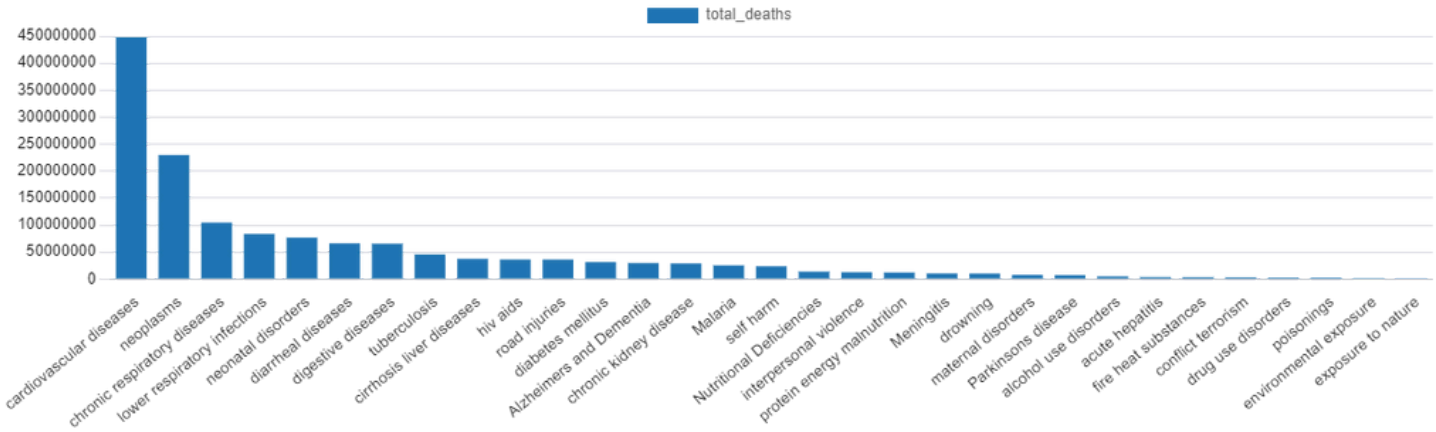
Data Output | Messages | Graph Visualiser

	disease text	total_deaths bigint
1	cardiovascular_diseases	447741982
2	neoplasms	229758538
3	chronic_respiratory_disea...	104605334
4	lower_respiratory_infectio...	83770038
5	neonatal_disorders	76860729
6	diarrheal_diseases	66235508
7	digestive_diseases	65638635
8	tuberculosis	45850603

Total 31 diseases have been queried and arranged according to highest to lowest deaths by a disease over all the years from 1990 to 2019.

A bar graph has been plotted below demonstrating total deaths per disease.

Result:- 3_diseases.csv



- Which year recorded the highest total deaths globally?

```sql
SELECT year, sum(total_deaths)
FROM cause_of_deaths
GROUP BY year
order by sum(total_deaths) desc
limit 1;
```

Data Output | Messages | Notifications

	year integer	sum bigint
1	2019	54362920

The highest number of deaths from the data of 30 years have been occurred in year 2019.

The total tally of deaths recorded in 2019 was about 54.36 million.

Result:- 4_highest_deaths_year.csv

- What is the trend in total deaths over time for all diseases?

```sql
SELECT year, sum(total_deaths)
FROM cause_of_deaths
GROUP BY year
ORDER BY year;
```

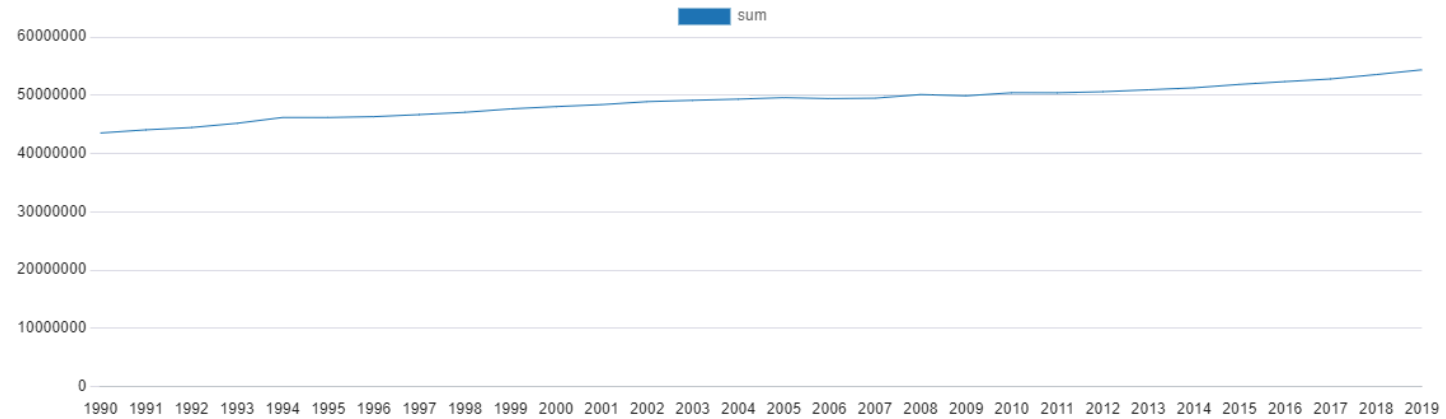
Data Output | Messages | Notifications

	year integer	sum bigint
1	1990	43518516
2	1991	44059729
3	1992	44459130
4	1993	45185713

The positive slope in the line graph plotted below demonstrates that the death toll have been always in an increasing order.

This may be due to various Socio-economic and Health-related factors.

Results:- 5_trend_of_deaths_over_years.csv



- Which country has the highest number of deaths from a cardiovascular diseases?

```
30  SELECT country_territory,
31  sum(cardiovascular_diseases)
32  as cardiovascular_diseases_deaths
33  FROM cause_of_deaths
34  GROUP BY country_territory
35  ORDER BY sum(cardiovascular_diseases) desc
36  LIMIT 1;
```

Data Output | Messages | Graph Visualiser | Notifications

	country_territory character varying (100)	cardiovascular_diseases_deaths bigint
1	China	100505973

As of 2019, from 1990s the highest number of deaths due to cardiovascular diseases occurred in China i.e. about 100 million.

This death toll may be due to various factors like use of tobacco, physical inactivity, unhealthy diet etc.

Results:-
6_cardiovascular_deaths.csv

- What is the total number of deaths per country over the entire dataset?

```
38  SELECT country_territory,
39  sum(total_deaths) as total_deaths
40  FROM cause_of_deaths
41  GROUP BY country_territory
42  Order BY sum(total_deaths) desc;
```

Data Output | Messages | Graph Visualiser | Notifica

	country_territory character varying (100)	total_deaths bigint
1	China	265408106
2	India	238158165
3	United States	71197802
4	Russia	59591155

Total rows: 204 | Query complete 00:00:00.243

The total number of deaths per country which data from includes 204 countries have been queried.

Results:- 7_deaths_per_country.csv

- How do death counts for Diabetes Mellitus vary between India, China and USA?

```sql
SELECT country_territory,
sum(diabetes_mellitus) as diabetes_mellitus_deaths
FROM cause_of_deaths
WHERE country_territory in ('India','China', 'United States')
GROUP BY country_territory
ORDER BY sum(diabetes_mellitus);
```

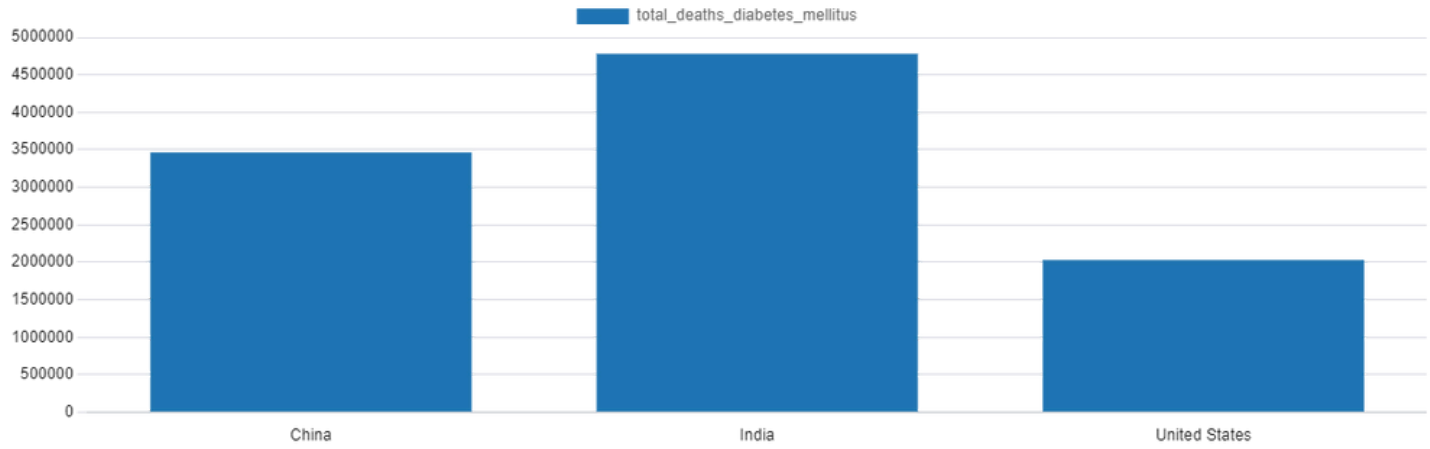
Data Output | Messages | Graph Visualiser | Notifications

	country_territory character varying (100)	diabetes_mellitus_deaths bigint
1	United States	2030631
2	China	3468554
3	India	4781169

Following are the death count by Diabetes Mellitus:

1. United states :- 2.03 million.
2. China :- 3.46 million.
3. India :- 4.78 million.

Results:-
8_total_deaths_by_diabetes_mellitus.csv



- What are the top 5 diseases causing the most deaths globally?

```sql
SELECT disease, total_deaths
from sum_and_union_of_diseases
ORDER BY total_deaths desc
LIMIT 5;
```

Data Output | Messages | Graph Visualiser

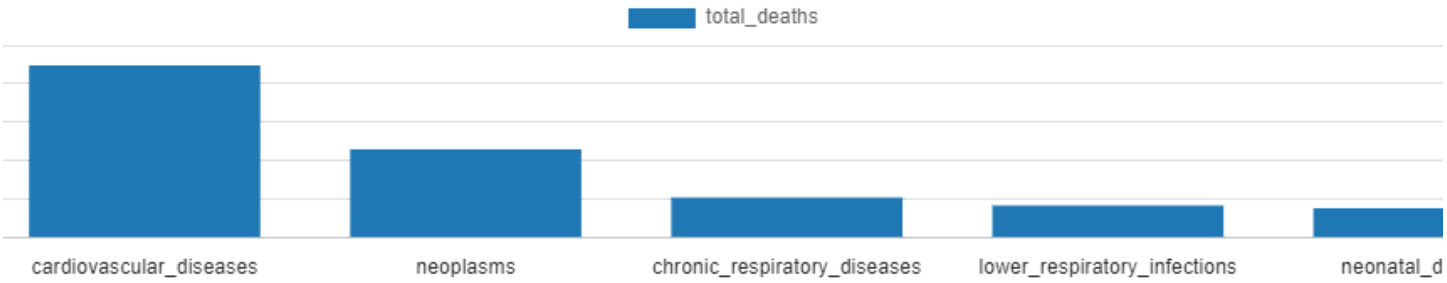
	disease text	total_deaths bigint
1	cardiovascular_diseases	447741982
2	neoplasms	229758538
3	chronic_respiratory_disea...	104605334
4	lower_respiratory_infections	83770038
5	neonatal_disorders	76860729

Total rows: 5 | Query complete 00:00:00.223

Top 5 diseases from the dataset which lead to death are:

1. Cardiovascular Diseases :- 447.74 million.
2. Neoplasm :- 229.75 million.
3. Chronic Respiratory Diseases :- 104.6 million.
4. Lower Respiratory Infections :- 83.77 million.
5. Neonatal_disorders :- 76.86 million.

Results:- 9_top_5_diseases.csv



- How have deaths from malaria changed over the years?

```
181  SELECT year,
182  sum(malaria) as deaths_by_malaria
183  FROM cause_of_deaths
184  GROUP BY year
185  ORDER BY year;
```

Data Output | Messages | Graph Visualiser | Notifi

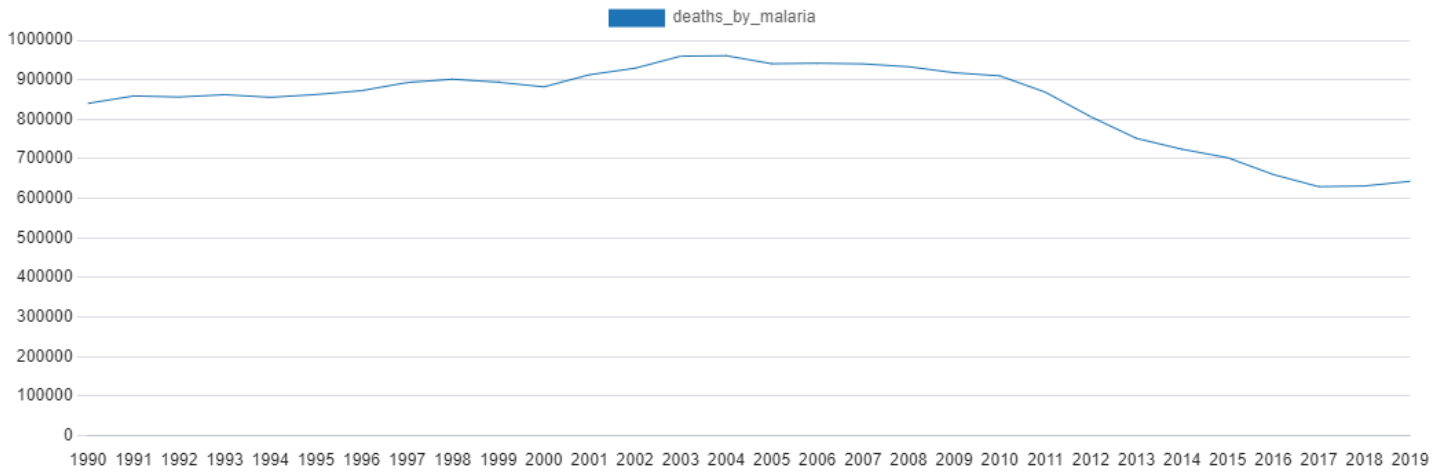
	year integer	deaths_by_malaria bigint
1	1990	840297
2	1991	858984
3	1992	856415
4	1993	862216
5	1994	855671

Total rows: 30 | Query complete 00:00:00.318

Line graph below demonstrates the change in death count by malaria over the years from 1990 to 2019 :

- There has been continuous increase in deaths by malaria from 1990 to 2004.The reasons may be due to lack of awareness, poor lifestyle and increased water pollution.
- After 2004 there has been a sharp fall in the count as shown in graph. This may be due to increase of awareness and improvement in healthcare quality.

Results:- 10_deaths_by_malaria.csv



- Which disease contributes to the most deaths in India?

```sql
SELECT disease,total_deaths
FROM (SELECT 'meningitis' AS disease, SUM(meningitis)
AS total_deaths FROM cause_of_deaths
WHERE country_territory = 'India'
UNION ALL
SELECT 'alzheimers_dementia', SUM(alzheimers_dementia
FROM cause_of_deaths WHERE country_territory = 'India
UNION ALL
/*(Repeat it with all disease Columns)*/
SELECT 'acute_hepatitis', SUM(acute_hepatitis)
FROM cause_of_deaths
WHERE country_territory = 'India')
AS disease_totals
ORDER BY total_deaths DESC
LIMIT 1;
```

Data Output | Messages | Graph Visualiser | Notifications

	disease text	total_deaths bigint
1	cardiovascular_diseas...	52994710

Highest number of deaths recorded in India from the dataset is about 53 million by cardiovascular diseases.

The main reasons for the high no. of deaths by cardiovascular diseases in India may be due to obesity, physical inactivity, unhealthy diet and increased tobacco use.

Results:- 11_India_highest_deaths.csv

- Which countries have the highest and lowest death rates for neonatal disorders?

```sql
(SELECT 'Highest' as category,
country_territory, sum(neonatal_disorders)
FROM cause_of_deaths
GROUP BY country_territory
ORDER BY  sum(neonatal_disorders) desc
LIMIT 1)
UNION ALL
(SELECT 'Lowest' as category,
country_territory, sum(neonatal_disorders)
FROM cause_of_deaths
GROUP BY country_territory
ORDER BY sum(neonatal_disorders)
LIMIT 1);
```

Data Output | Messages | Graph Visualiser | Notifications

	category text	country_territory character varying (100)	sum bigint
1	Highest	India	20911570
2	Lowest	Niue	1

Highest and Lowest neonatal deaths occurred in India and Niue respectively.

High neonatal deaths in India may be due to poor life quality, cultural barriers, increases pollution also.

Results:- 12_neonatal_deaths.csv

- How many deaths have been occurred by road injuries in India to that of China?

```sql
SELECT country_territory,
sum(road_injuries) as road_injury_deaths
FROM cause_of_deaths
WHERE country_territory in ('India', 'China')
GROUP BY country_territory
ORDER BY sum(road_injuries) desc;
```

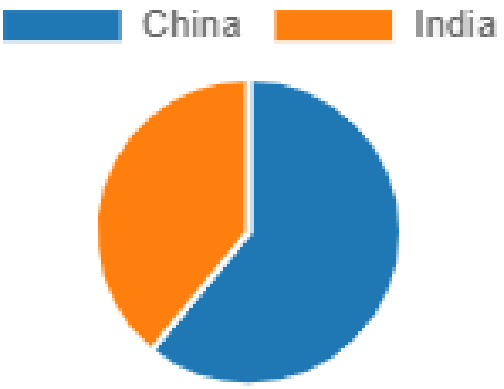
Data Output | Messages | Graph Visualiser | Notifications

	country_territory character varying (100)	road_injury_deaths bigint
1	China	8350399
2	India	5346154

Number of deaths by road injuries in India is about 5.34 million which is less than that of China with 83.5 million road injury deaths.

China was more populous country than India over all the years, which is one of the main reasons why the deaths by road injuries are greater that of India.

Results:- 13_road_deaths.csv



- which year have resulted in highest and lowest deaths by Tuberculosis?

```sql
(SELECT 'Highest' as category,
year, sum(tuberculosis)
FROM cause_of_deaths
GROUP BY year
ORDER BY  sum(tuberculosis) desc
LIMIT 1)
UNION ALL
(SELECT 'Lowest' as category,
year, sum(tuberculosis)
FROM cause_of_deaths
GROUP BY year
ORDER BY sum(tuberculosis)
LIMIT 1);
```

Data Output | Messages | Graph Visualiser | Nc

	category text	year integer	sum bigint
1	Highest	1992	1807740
2	Lowest	2019	1179234

Highest no. of deaths by tuberculosis occurred in year 1992. Reasons for high death count was overcrowding, poor sanitation and lack of quality drugs.

Lowest count of death by tuberculosis was noted in 2019 . This was possible due to new innovations in healthcare technologies, drug quality and good sanitation in developed countries. Although there is need of reducing the count in developing and under-developed countries.

Results:- 14_tuberculosis_high_low.csv

- Compare trend of deaths between Alzihemers dimentia and Parkinsons disease over all the years

```
11  SELECT year, sum(alzheimers_dementia)
12  as alzheimers_dementia_deaths,
13  sum(parkinsons) as parkinsons_deaths
14  FROM cause_of_deaths
15  GROUP BY year
16  ORDER BY year;
```

Data Output | Messages | Graph Visualiser | Notifications

	year integer	alzheimers_dementia_deaths bigint	parkinsons_deaths bigint
1	1990	560616	147156
2	1991	583166	150875
3	1992	605894	154886
4	1993	629571	160249
5	1994	652176	164381
6	1995	674815	168882

Total rows: 30 | Query complete 00:00:00.330

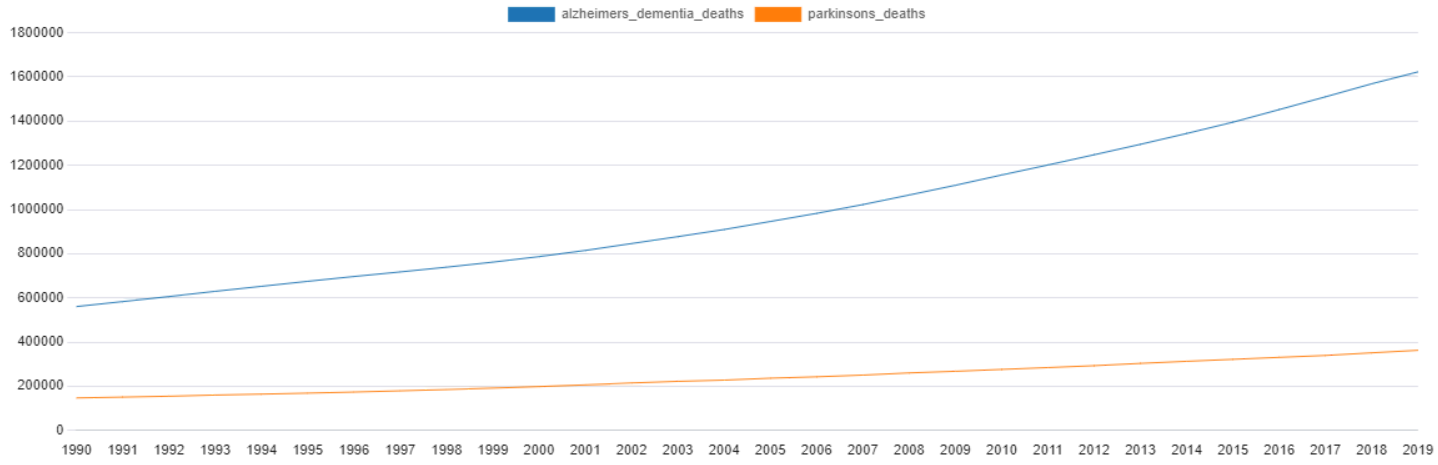
The count of deaths by Alzihemers dimentia and Parkinsons have been queried and visualized in below line graph.

Deaths by Alzihemers and dimentia have been continuously in a positive trend. Main reason for it may be poor life quality.

Deaths by Parkinsons has also been in a positive trend always. This may be due to heart diseases and pneumonia.

Results:-

15_deaths_by_alzheimer_parkinsins.csv



Prescriptive Analysis:

Insights Derived from Data Analysis

- High-Risk Diseases:
  - Cardiovascular diseases are the leading cause of death globally, with China recording the highest deaths.
  - Neonatal disorders have significantly high mortality rates, particularly in India.
  - Road injuries remain a major concern, with China experiencing the highest fatalities.
- Temporal Trends:
  - The overall number of deaths has been increasing from 1990 to 2019.

    - Malaria deaths increased until 2004 but declined afterward due to improved healthcare measures.
    - Tuberculosis deaths peaked in 1992 and declined over time, suggesting better sanitation and medical advancements.
- Regional Variations:
    - Developed countries have lower mortality rates due to better healthcare systems.
    - Developing and underdeveloped countries struggle with neonatal and infectious diseases due to inadequate medical facilities.
- Specific Disease Trends:
    - Alzheimer's and Parkinson's diseases show a continuous increase in deaths, likely due to aging populations and lifestyle factors.
    - Diabetes-related deaths are highest in India, followed by China and the USA, indicating concerns over diet and physical inactivity.

Preventive Measures

- Cardiovascular Disease Prevention:
    - Promote heart-healthy diets, regular physical activity, and smoking cessation programs.
    - Implement nationwide screening for hypertension and diabetes.
- Neonatal Disorder Reduction:
    - Improve prenatal and postnatal care in developing countries.
    - Raise awareness about maternal nutrition and infant healthcare.
- Road Injury Mitigation:
    - Strengthen traffic regulations and enforcement in high-risk regions.
    - Improve road infrastructure and promote vehicle safety measures.
- Infectious Disease Control:
    - Increase vaccination efforts, particularly for malaria, tuberculosis, and neonatal infections.
    - Ensure access to clean drinking water and sanitation facilities.

Recommended Steps for Stakeholders

- Governments & Policymakers:
    - Allocate healthcare resources based on mortality trends.
    - Develop targeted intervention programs for high-risk diseases.
- Healthcare Institutions:
    - Enhance diagnostic and treatment facilities for cardiovascular diseases, diabetes, and infectious diseases.
    - Invest in AI-driven predictive healthcare solutions to forecast disease outbreaks.
- Public Health Organizations:
    - Conduct awareness campaigns on preventable diseases.
    - Encourage lifestyle changes through community engagement programs.

Recommended Steps for the Public

- Heart Health – Eat healthy, exercise regularly, avoid smoking, and monitor BP/cholesterol.
- Diabetes Prevention – Maintain a balanced diet, stay active, and monitor blood sugar levels.
- Road Safety – Wear seat belts/helmets, follow traffic rules, and avoid distractions while driving.
- Infectious Disease Control – Maintain hygiene, drink clean water, and get vaccinated.
- Neonatal Care – Pregnant women should have regular check-ups, proper nutrition, and newborn vaccinations.
- Mental Health & Cognitive Wellness – Stay mentally active, exercise, and seek help when needed.
- Community Awareness – Educate others, participate in health campaigns, and promote hygiene.

Conclusion

The analysis highlights the urgent need for targeted health interventions, particularly in developing countries. A data-driven approach is essential for effective healthcare planning, ensuring that resources are optimally utilized. By implementing the prescribed measures, stakeholders can significantly reduce preventable deaths and improve global public health outcomes.