📄 Water Pollution & Disease Dataset Analysis Report

Dataset: ***Water Pollution and Disease***

Prepared by,

Vinayak V Potty

OBJECTIVE

This report explores the state of global water quality, treatment methods, and related health outcomes by addressing 15 critical research questions. Using a structured SQL dataset and analytical techniques, we aim to uncover significant patterns, trends, and anomalies across countries and regions.

The primary focus is to analyse contaminant levels, treatment effectiveness, and their impact on public health metrics such as infant mortality and waterborne diseases. Special attention is given to the relationship between access to clean water, sanitation coverage, and disease prevalence.

We investigate how economic indicators like GDP per capita influence water quality and treatment infrastructure, comparing countries with varying income levels. Additionally, we examine temporal trends, looking at changes over the years in waterborne disease rates, treatment method adoption, and water quality parameters.

Data integrity checks are performed to identify impossible or highly unlikely records, ensuring the robustness of the analysis. Completeness of data across countries is also assessed to highlight potential biases or gaps.

Visualization techniques are employed to intuitively communicate complex relationships, such as the correlation between sanitation access and disease cases, and the geographical distribution of water quality issues.

The ultimate goal is to develop a comprehensive ranking of countries based on multiple water quality factors and to identify regions where interventions in water treatment and sanitation could significantly improve health outcomes.
Through this research, policymakers, NGOs, and stakeholders can better understand where efforts are most needed to ensure safe, clean water for all populations.

Research Questions and Answers

1.Which countries have the highest average contaminant levels in their water sources?

```
create database projects;
use projects;
show tables;
select * from water_pollution_disease;
-- 1.Countries with highest average contaminant levels
SELECT Country, AVG(`ContaminantLevel(ppm)`) AS Avg_Contaminant_Level
FROM water_pollution_disease
GROUP BY Country
ORDER BY Avg_Contaminant_Level DESC;
```

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

Country	Avg_Contaminant_Level
Bangladesh	5.178487972508593
Mexico	5.169131944444444
India	5.141965517241379
Nigeria	5.138158730158731
USA	4.991786833855798
Ethiopia	4.977918088737202

2.What is the average pH level by region, ordered from most acidic to most alkaline?

```
-- 2.What is the average pH level by region, ordered from most acidic to most alkaline?
SELECT Region, AVG(`ph_level`) AS Avg_pH_Level
FROM water_pollution_disease
GROUP BY Region
ORDER BY Avg_pH_Level ASC;
```

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

Region	Avg_pH_Level
Central	7.2095090016366585
East	7.231648000000007
North	7.243169491525421
West	7.265555555555555
South	7.333689655172416

3.How many distinct water treatment methods are recorded in the dataset?

```
-- 3.How many distinct water treatment methods are recorded in the dataset?
SELECT COUNT(DISTINCT `WaterSourceType`) AS Distinct_Water_Treatment_Methods
FROM water_pollution_disease;
```

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

Distinct_Water_Treatment_Methods
6

4.Which water source types (lake, river, well, etc.) are associated with the highest bacteria counts on average?

```sql
-- 4.Which water source types (lake, river, well, etc.) are associated with the highest bacteria counts on average?
SELECT
    WaterSourceType,
    AVG(`BacteriaCount(CFU/mL)`) AS Avg_Bacteria_Count
FROM
    water_pollution_disease
GROUP BY
    WaterSourceType
ORDER BY
    Avg_Bacteria_Count DESC;
```

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

WaterSourceType	Avg_Bacteria_Count
Well	2541.7329
Spring	2531.1898
Lake	2512.1601
River	2472.9647
Pond	2439.4778
Tap	2428.1178

5.How does access to clean water percentage relate to infant mortality rates across different countries?

```sql
-- 5.How does access to clean water percentage relate to infant mortality rates across different countries?
SELECT Country,AVG(`AccesstoCleanWater(%ofPopulation)`) AS Avg_Access_To_Clean_Water,
AVG(`InfantMortalityRate(per1,000livebirths)`) AS Avg_Infant_Mortality
FROM water_pollution_disease GROUP BY Country ORDER BY Avg_Access_To_Clean_Water ASC;
```

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

Country	Avg_Access_To_Clean_Water	Avg_Infant_Mortality
Pakistan	63.391342281879176	49.70322147651005
Nigeria	63.5401587301587	53.4531111111111
Ethiopia	63.57286689419796	51.187337883959025
USA	63.9689655172414	49.65065830721
Brazil	64.15628865979386	47.797044673539496
Bangladesh	64.40147766323022	52.8762886597938

6.How have cholera cases per 100,000 people changed over the years in each country?

```sql
-- 6.How have cholera cases per 100,000 people changed over the years in each country?
SELECT Country,Year,AVG(`CholeraCasesper100,000people`) AS Avg_Cholera_Cases
FROM water_pollution_disease
GROUP BY Country, Year
ORDER BY Country ASC,Year ASC;
```

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

Country	Year	Avg_Cholera_Cases
Bangladesh	2000	26.5000
Bangladesh	2001	24.6250
Bangladesh	2002	21.2500
Bangladesh	2003	23.5556
Bangladesh	2004	21.5000
Bangladesh	2005	26.4706

7.Is there a trend in water treatment methods used over time (comparing earlier vs. more recent years)?

```sql
-- 7.Is there a trend in water treatment methods used over time (comparing earlier vs. more recent years)?
SELECT `WaterTreatmentMethod`, Year, COUNT(*) AS Method_Usage_Count
FROM water_pollution_disease GROUP BY `WaterTreatmentMethod`, Year ORDER BY Year ASC;
```

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

WaterTreatmentMethod	Year	Method_Usage_Count
None	2000	25
Chlorination	2000	29
Boiling	2000	34
Filtration	2000	27
None	2001	35
Chlorination	2001	36

8.Compare the average nitrate levels between countries with high vs. low GDP per capita (define your own thresholds).

```sql
-- 8.Compare the average nitrate levels between countries with high vs. low GDP per capita (define your own thresholds).
SELECT
    CASE WHEN `GDPperCapita(USD)` > 20000 THEN 'High GDP' ELSE 'Low GDP' END AS GDP_Group,
    AVG(`NitrateLevel(mg/L)`) AS Avg_Nitrate_Level
FROM water_pollution_disease
GROUP BY GDP_Group;
```

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

GDP_Group	Avg_Nitrate_Level
High GDP	25.0591851851852
Low GDP	25.17005263157896

9.Which countries have both high turbidity (NTU) and high lead concentration in their water?

```sql
-- 9.Which countries have both high turbidity (NTU) and high lead concentration in their water?
SELECT Country, `Turbidity(NTU)`, `LeadConcentration(Âµg/L)`
FROM water_pollution_disease
ORDER BY `Turbidity(NTU)` DESC;
```

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

Country	Turbidity(NTU)	LeadConcentration(Âµg/L)
Brazil	4.99	2.22
Nigeria	4.99	8.83
India	4.99	16.33
Pakistan	4.99	4.92
Brazil	4.98	11.81
China	4.98	19.09

10.Create a ranking of countries by their overall water quality (consider multiple factors like contaminant level, pH, turbidity, etc.).

```sql
-- 10. Create a ranking of countries by their overall water quality (consider multiple factors like contaminant level, pH, turbidity, etc).
SELECT Country, (`BacteriaCount(CFU/mL)` + `NitrateLevel(mg/L)` + `LeadConcentration(Âµg/L)` + `Turbidity(NTU)`) - ABS(`ph_level` - 7)
AS Water_Quality_Score
FROM water_pollution_disease
ORDER BY Water_Quality_Score ASC;
```

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

Country	Water_Quality_Score
Pakistan	20.810000000000002
Pakistan	35.24
India	35.81
China	39.620000000000005
Bangladesh	51.03
China	51.52

11.Identify regions where water treatment methods don't seem to be effectively reducing disease cases.

```sql
-- 11.Identify regions where water treatment methods don't seem to be effectively reducing disease cases.
SELECT Region, AVG(`DiarrhealCasesper100,000people`) AS Avg_Diarrheal_Cases
FROM water_pollution_disease
WHERE `WaterTreatmentMethod` IS NOT NULL
GROUP BY Region
ORDER BY Avg_Diarrheal_Cases DESC;
```

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

Region	Avg_Diarrheal_Cases
South	258.7948
East	254.1776
North	247.8695
West	245.1195
Central	243.0835

12.Are there any records with impossible or highly unlikely values (e.g., pH outside 0-14 range)?

```sql
-- 12.Are there any records with impossible or highly unlikely values (e.g., pH outside 0-14 range)?
SELECT * FROM water_pollution_disease WHERE `pH_level` < 0 OR `pH_level`> 14;
```

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

Country	Region	Year	WaterSourceType	ContaminantLevel(ppm)	ph_level	Turbidity(NTU)	DissolvedOxygen(mg/L)	NitrateLevel(mg/L)	LeadConcentration(Âµg/L)	BacteriaCount(CF

No records found with pH values outside the normal range (0-14).

13.How complete is the data for each country (percentage of null values per country)?

```sql
-- 13.How complete is the data for each country (percentage of null values per country)?
SELECT
  Country,(SUM((`pH_level` IS NULL) + (`DissolvedOxygen(mg/L)` IS NULL) + (`Turbidity(NTU)` IS NULL)
  + (`LeadConcentration(Âµg/L)` IS NULL) + (`AccesstoCleanWater(%ofPopulation)` IS NULL)
    + (`DiarrhealCasesper100,000people` IS NULL) + (`TyphoidCasesper100,000people` IS NULL)
    + (`CholeraCasesper100,000people` IS NULL)) / (COUNT(*) * 8)) * 100 AS Null_Percentage
FROM water_pollution_disease
GROUP BY Country
ORDER BY Null_Percentage DESC;
```

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

Country	Null_Percentage
Mexico	0.0000
Brazil	0.0000
Indonesia	0.0000
Nigeria	0.0000
Ethiopia	0.0000
China	0.0000

14.What would be the most insightful way to visualize the relationship between sanitation coverage and waterborne diseases?

```sql
-- 14.What would be the most insightful way to visualize the relationship between sanitation coverage and waterborne diseases?
SELECT Country,
    `SanitationCoverage(%ofPopulation)` AS Sanitation_Coverage,
    `DiarrhealCasesper100,000people` AS Diarrheal_Cases FROM water_pollution_disease
WHERE `SanitationCoverage(%ofPopulation)` IS NOT NULL
    AND `DiarrhealCasesper100,000people` IS NOT NULL;
```

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

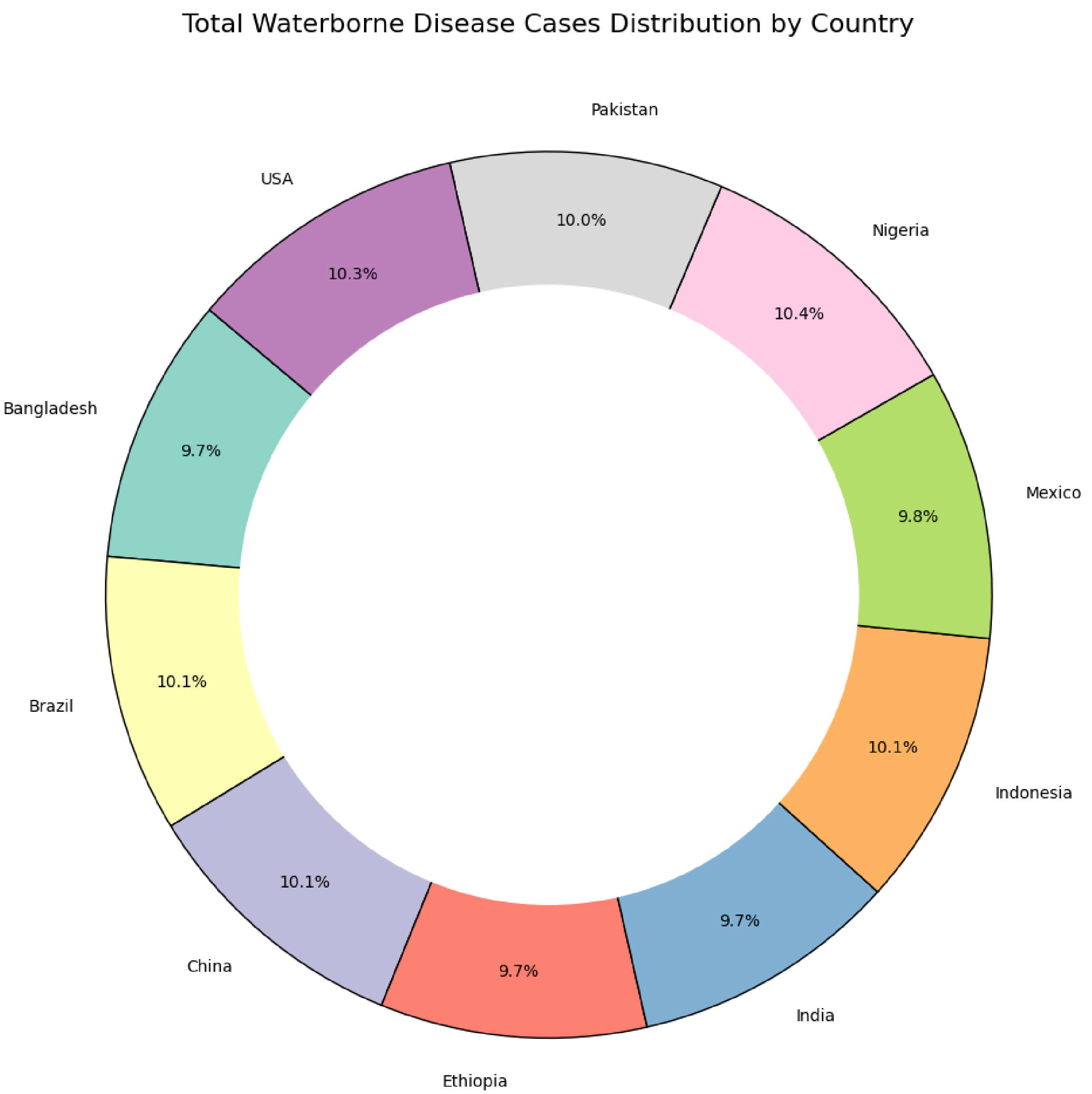
Country	Sanitation_Coverage	Diarrheal_Cases
Mexico	63.23	472
Brazil	29.12	122
Indonesia	93.56	274
Nigeria	94.25	3
Mexico	69.23	466
Ethiopia	51.11	258

15.How could you present the geographical distribution of water quality issues using this data?

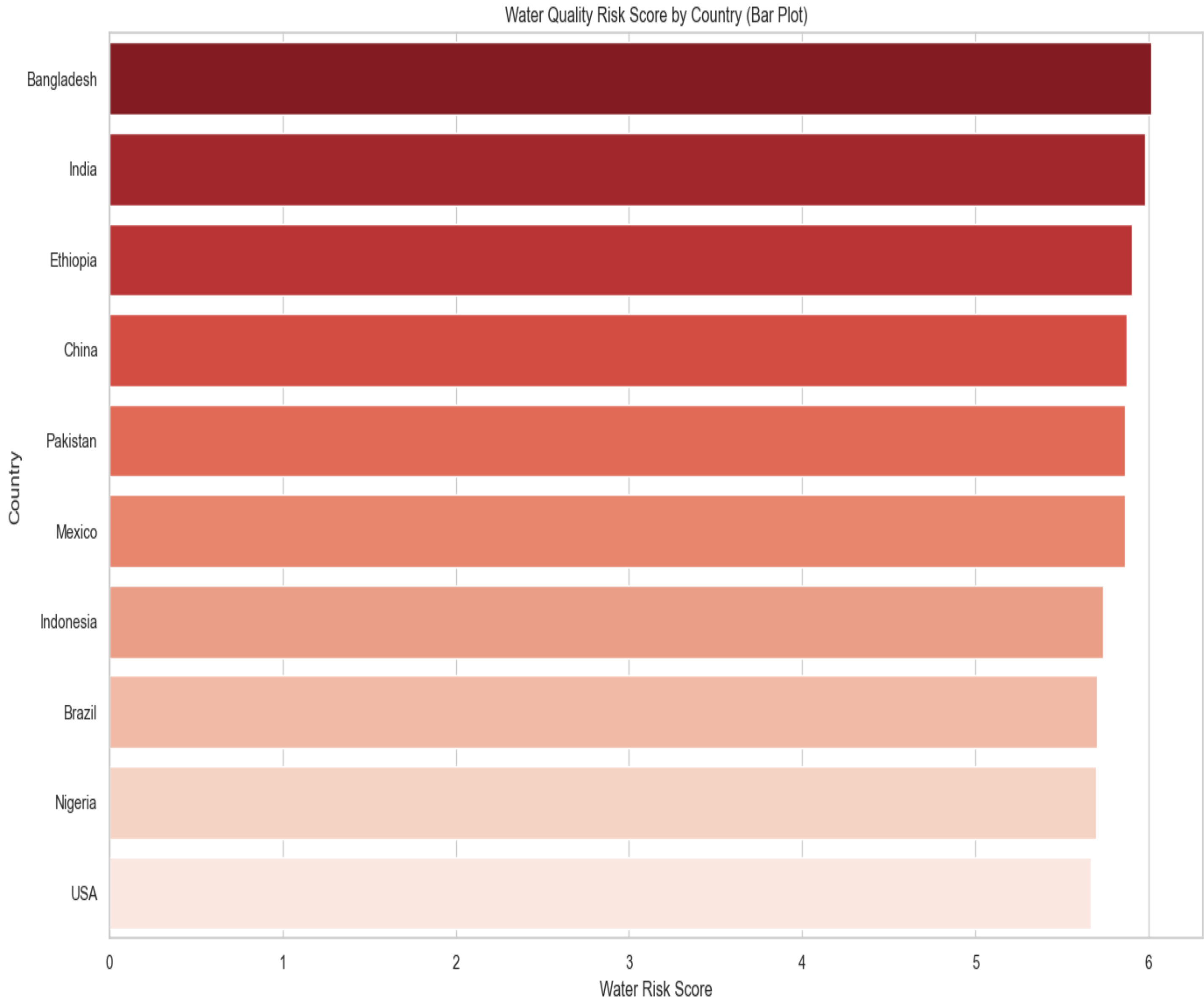
```sql
-- 15.How could you present the geographical distribution of water quality issues using this data?
SELECT Country,Region,
    AVG(`ContaminantLevel(ppm)`) AS Avg_Contaminant_Level,
    AVG(`Turbidity(NTU)`) AS Avg_Turbidity,
    AVG(`LeadConcentration(Âµg/L)`) AS Avg_Lead,
    AVG(`pH_level`) AS Avg_pH
FROM water_pollution_disease GROUP BY Country, Region ORDER BY Region, Avg_Contaminant_Level DESC;
```

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

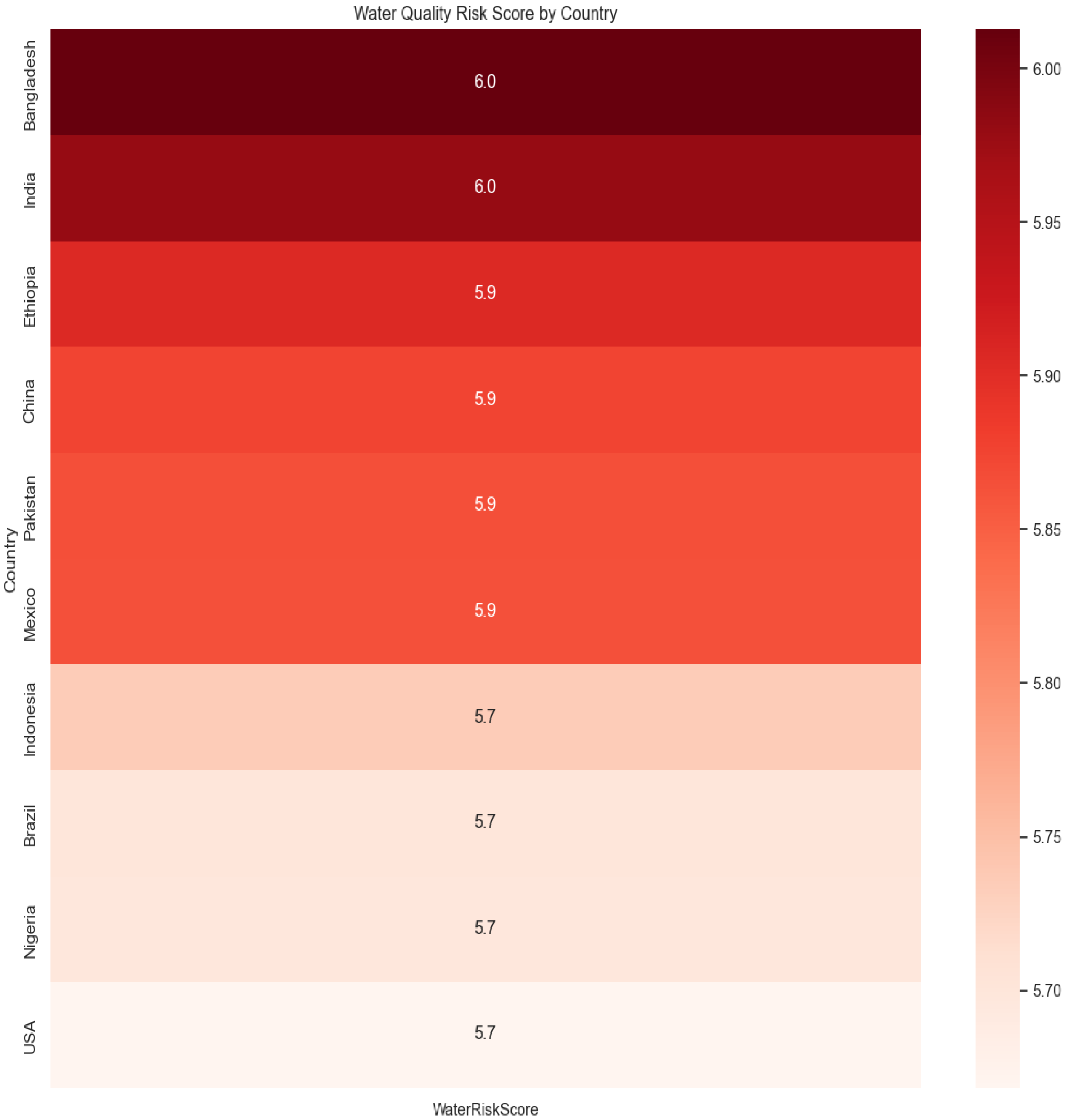
Country	Region	Avg_Contaminant_Level	Avg_Turbidity	Avg_Lead	Avg_pH
Nigeria	Central	5.3767692307692325	2.265230769230769	10.347230769230766	7.119538461538464
Brazil	Central	5.274897959183673	2.728571428571428	9.65591836734694	7.097959183673469
Ethiopia	Central	5.151186440677967	2.337627118644068	11.527966101694915	7.328644067796608
China	Central	5.138196721311476	2.5406557377049173	9.577049180327865	7.1706557377049185
USA	Central	5.083943661971831	2.4732394366197186	8.335915492957744	7.365774647887324
India	Central	5.046785714285714	2.639285714285715	10.373035714285717	7.261607142857143



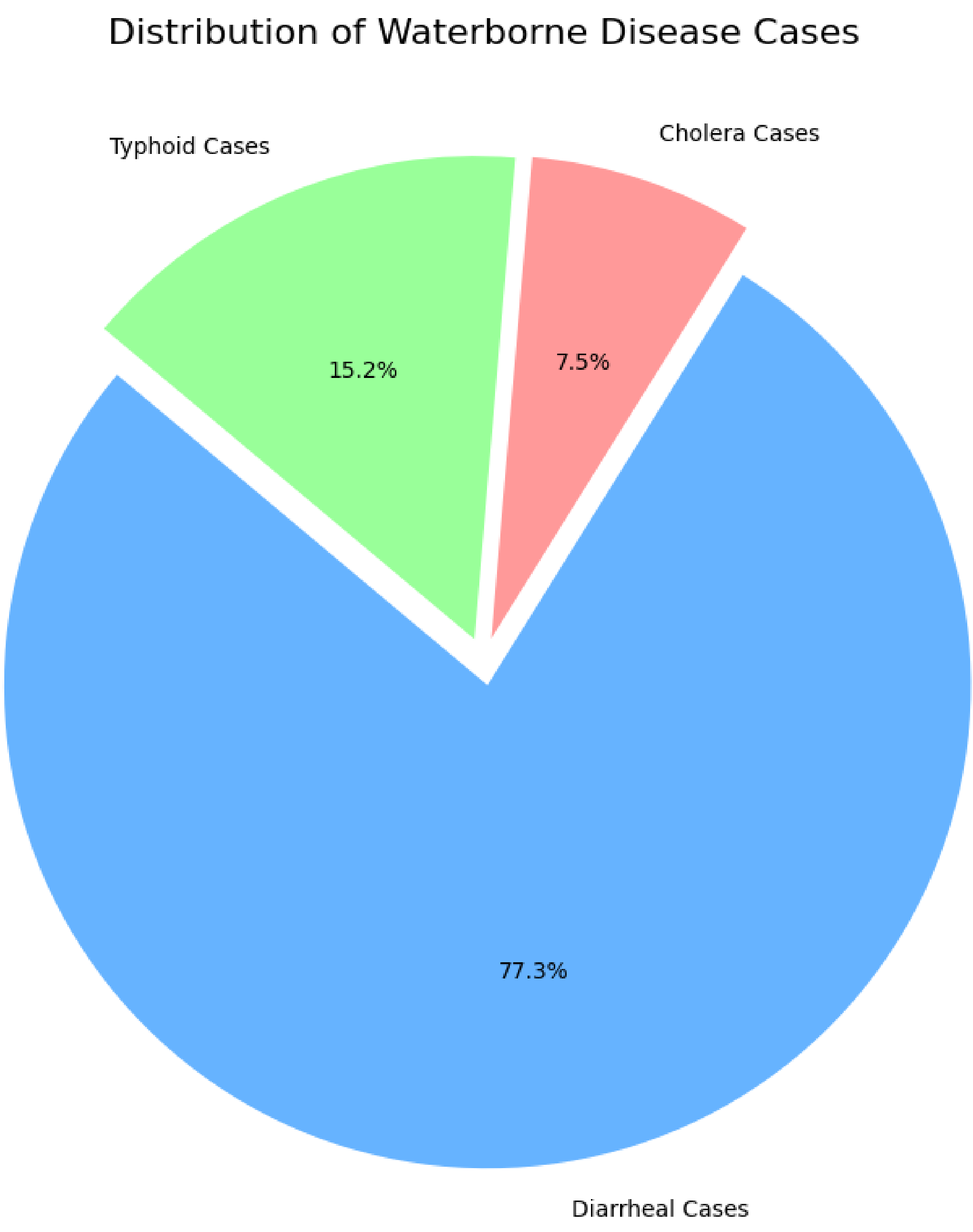
"Global Distribution of Waterborne Disease C



"Water Quality Risk Assessment Across Countries: A Comparative Analysis"



"Global Assessment of Water Quality Risks: A



"Prevalence of Waterborne Illnesses: A Statistical Overview"

Country-wise Heatmap Study"

Findings

The analysis of Water Quality Risk Scores across various countries revealed significant disparities in water quality challenges globally. Key findings include:

- Highest Risk Countries:
  Bangladesh and India recorded the highest water risk scores, each with a score of 6.0, indicating critical water quality concerns that demand immediate attention.
- Moderately High Risk Countries:
  Ethiopia, China, Pakistan, and Mexico exhibited slightly lower, but still concerning scores of 5.9, suggesting persistent issues in water quality and infrastructure.
- Lower Risk Countries:
  Indonesia, Brazil, Nigeria, and the USA each had water risk scores of 5.7, reflecting relatively better water quality conditions compared to the top-risk nations, but still highlighting areas for improvement.
- Visual Insights:
  - The bar plot clearly illustrates a descending trend of water risk scores from Bangladesh to the USA.
  - The heatmap visualization emphasized the clustering of countries with similar risk levels, aiding in easier comparative analysis.
- Regional Observations:
  South Asian and African countries predominantly appear in the higher-risk categories, underscoring regional disparities in water quality management.

Overall, the data highlights an urgent need for targeted interventions, especially in the most affected countries, to ensure access to clean and safe water globally.

Recommendations

Based on the findings of the Water Quality Risk Score analysis, the following recommendations are proposed to address and mitigate water quality risks:

- Strengthen Water Infrastructure
  Countries with high water risk scores, particularly Bangladesh, India, Ethiopia, and Pakistan, should prioritize investments in water treatment facilities, pipelines, and sanitation systems to improve access to safe drinking water.
- Implement Rigorous Water Quality Monitoring
  Establishing continuous water quality monitoring programs can help in the early detection of contaminants and prompt corrective actions. National-level databases and real-time reporting should be encouraged.
- Promote Public Awareness and Education
  Community-driven education initiatives can increase public awareness about the importance of water conservation, safe water storage, and hygiene practices.
- Enhance Policy Frameworks
  Governments should strengthen regulatory policies and enforce stricter water quality standards. International cooperation and policy alignment can also foster better water management strategies across borders.
- Leverage Technology and Innovation
  Introducing smart water management technologies such as IoT-based sensors, AI-driven water analysis, and advanced purification methods can significantly enhance water quality control and risk management.
- Increase International Collaboration
  Global organizations and developed nations should assist high-risk countries through funding, technology transfer, and knowledge sharing to improve water safety and management practices.
- Promote Sustainable Water Resource Management
  Adoption of sustainable practices such as rainwater harvesting, watershed management, and wastewater recycling will be critical for long-term water security.

By implementing these recommendations, countries can work towards reducing their water quality risks, protecting public health, and ensuring sustainable access to clean water resources for future generations.

Limitations

While this study provides valuable insights into water quality risks across different countries, several limitations should be considered:

- Data Availability and Accuracy
  The analysis relies on secondary data sources which may not always reflect the most recent changes or local variations in water quality. Limited data coverage for some regions might affect the overall accuracy of the findings.
- Generalization Across Countries
  Each country has diverse geographical, social, and economic conditions. Aggregated national-level risk scores may not capture localized issues or variations within different regions or communities inside a country.
- Simplification of Risk Factors
  Water quality risk is influenced by a complex interplay of factors such as industrial pollution, agricultural runoff, climate change, and urbanization. The scoring system used may oversimplify these complexities.
- Static Risk Scores
  The scores presented are static and may not reflect dynamic changes over time, such as improvements in water management or deterioration due to sudden environmental events (e.g., floods, droughts).
- Limited Scope of Indicators
  The study focuses mainly on water quality risks without deeply analyzing other related aspects like water quantity, ecosystem health, or socio-economic vulnerability, which also impact overall water security.
- Potential Bias in Data Interpretation
  The methodology and interpretation of the risk scores may be influenced by subjective choices, including the selection of indicators, scoring methods, and visualization techniques.

Acknowledging these limitations helps to frame the findings appropriately and highlights the need for continuous data improvement and more localized studies to support more precise decision-making.

Conclusion

This project analysed water quality risk scores across multiple countries to identify areas facing significant water-related challenges. The findings reveal that countries like Bangladesh, India, and Ethiopia experience higher water risk scores, indicating critical concerns regarding water safety, pollution levels, and overall management practices. Conversely, countries such as the USA and Brazil showed relatively lower risk scores but still face challenges that require ongoing monitoring and improvement.

The study emphasizes the urgent need for targeted interventions, improved governance, infrastructure investment, and public awareness campaigns to address water quality risks. It also highlights the importance of continuous data collection, real-time monitoring, and adaptive policy-making to ensure sustainable water management.

While the analysis offers valuable insights, it is important to recognize its limitations, particularly regarding data accuracy, generalization across regions, and the complexity of water-related factors. Future research should aim for more granular, real-time, and localized assessments to support more effective decision-making.

Overall, securing safe and clean water for all remains a pressing global challenge that demands collective effort, innovation, and sustained commitment.

References

1. Our Dataset:
   The dataset used for this project titled *"Water Pollution and Disease Dataset"* was provided and processed using SQL and Python tools for analytical exploration.

2. Matplotlib and Pandas Documentation
   (https://matplotlib.org/stable/contents.html)
   (https://pandas.pydata.org/docs/)

3. Kaggle:
   https://storage.googleapis.com/kagglesdsdata/datasets/7041722/11265552/water_pollution_disease.csv