Practical -9,10: Mini Project – Phase 1&2

Solution:

Dataset: “water_pollution_disease.csv”

The dataset consists of **3000 rows** and **24 columns**, capturing data from various countries and regions across different years. It includes water quality indicators, health statistics, and socio-economic factors, providing insights into the relationship between environmental conditions and public health.

Output:

```
df.head(10)
```

	Country	Region	Year	Water Source Type	Contaminant Level (ppm)	pH Level	Turbidity (NTU)	Dissolved Oxygen (mg/L)	Nitrate Level (mg/L)	Lead Concentration (µg/L)	...	GDP per Capita (USD)	Healthcare Access Index (0-100)	Urbanization Rate (%)	Sanitation Coverage (% of Population)	Rainfall (mm per year)	Temperature (°C)	Population Density (people per km²)	Year Range	Health Score	Water Quality Score
0	Mexico	North	2015	Lake	6.06	7.12	3.93	4.28	8.28	7.89	...	57057	96.92	84.61	63.23	2800	4.94	593	2015	625.16	483.080000
1	Brazil	West	2017	Well	5.24	7.84	4.79	3.86	15.74	14.68	...	17220	84.73	73.37	29.12	1572	16.93	234	2015	234.30	310.592857
2	Indonesia	Central	2022	Pond	0.24	6.43	0.79	3.42	36.67	9.96	...	86022	58.37	72.86	93.56	2074	21.73	57	2020	411.45	341.072857
3	Nigeria	East	2016	Well	7.91	6.71	1.96	3.12	36.92	6.77	...	31166	39.07	71.07	94.25	937	3.79	555	2015	144.66	548.912857
4	Mexico	South	2005	Well	0.12	8.16	4.22	9.15	49.35	12.51	...	25661	23.03	55.55	69.23	2295	31.44	414	2005	623.78	609.358571
5	Ethiopia	West	2013	Tap	2.93	8.21	4.03	8.66	31.35	16.74	...	84334	53.45	86.11	51.11	2530	8.01	775	2010	405.13	135.988571
6	China	East	2022	River	0.06	6.11	3.12	6.97	44.98	1.23	...	6726	34.56	76.53	68.60	1573	22.42	584	2020	413.59	434.210000
7	Mexico	Central	2024	Spring	3.76	6.42	1.35	9.99	2.73	19.44	...	76593	7.27	45.22	70.22	940	19.64	111	2020	474.30	173.670000
8	Indonesia	North	2014	Well	0.63	6.29	1.42	9.67	28.31	1.93	...	5470	47.62	69.40	69.77	2150	36.23	538	2010	361.32	29.607143
9	Nigeria	Central	2013	Well	9.14	6.45	0.62	7.59	45.99	10.23	...	72858	57.11	89.85	78.67	2083	35.73	250	2010	305.43	367.574286

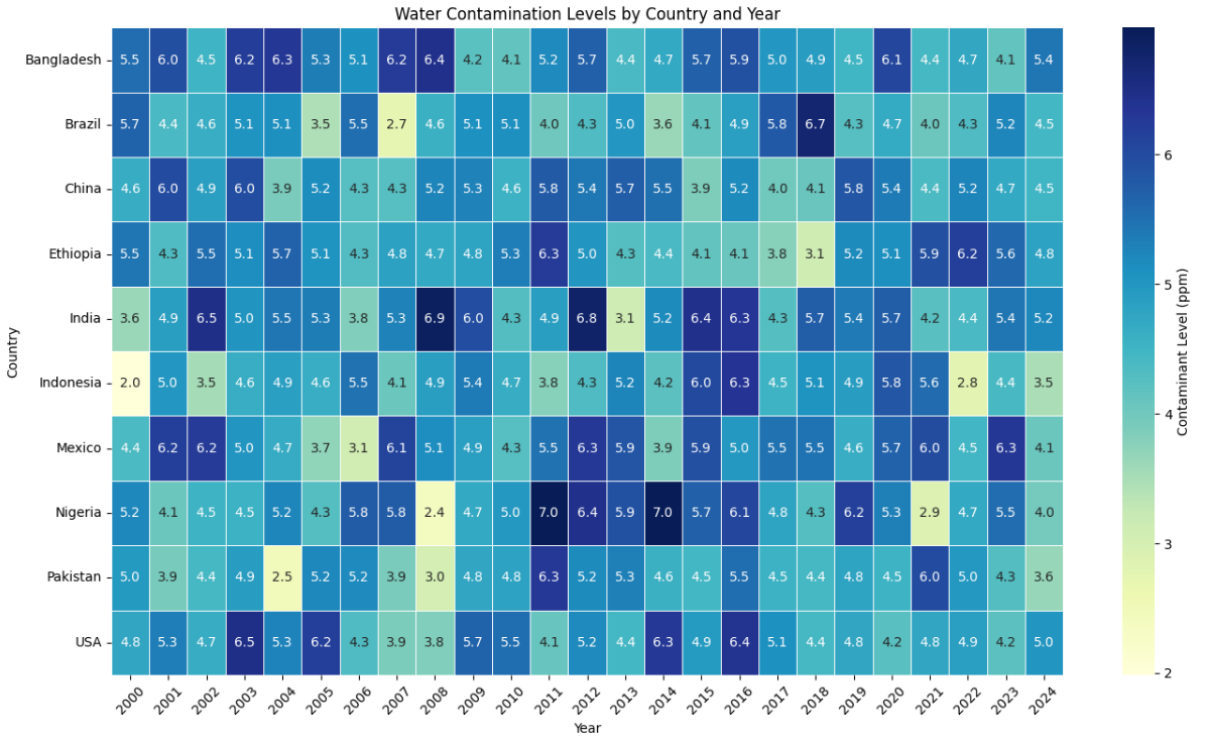
10 rows × 27 columns

DESCRIPTIVE STATISTICS

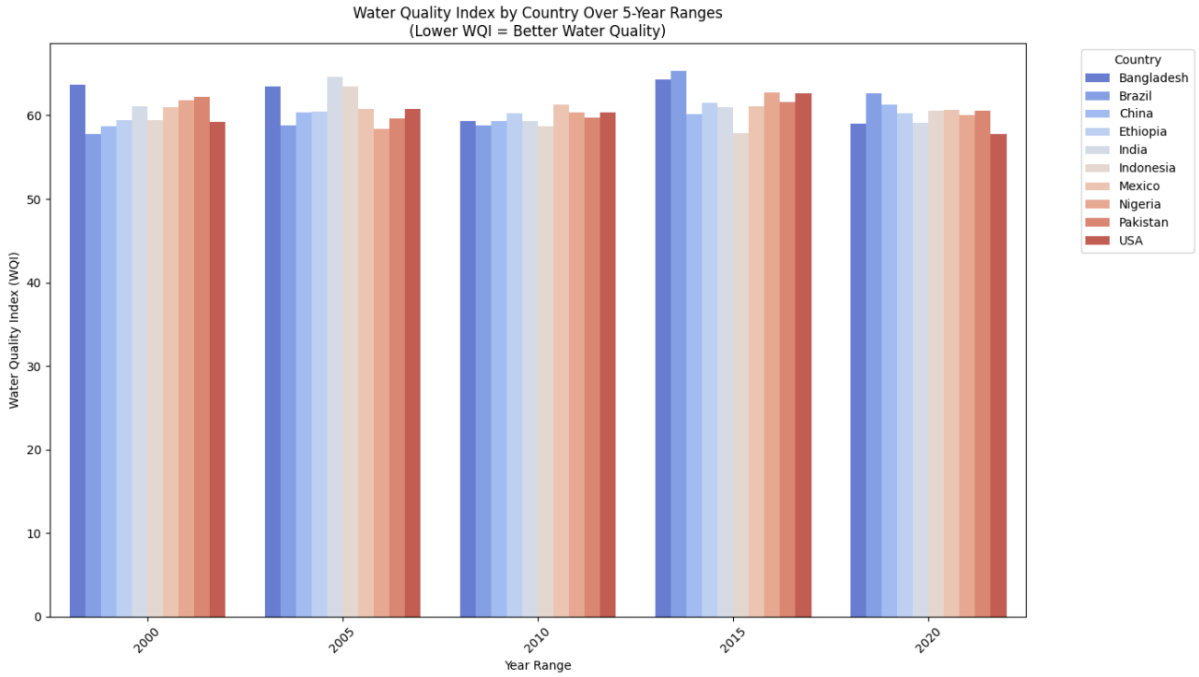
	Mean	Median	Mode	Standard Deviation	Variance	Range	IQR
Year	2012.01	2012.00	2006.00	7.23	5.226000e+01	24.00	12.00
Contaminant Level (ppm)	4.95	4.95	2.56	2.86	8.180000e+00	10.00	4.84
pH Level	7.26	7.28	7.08	0.72	5.200000e-01	2.50	1.24
Turbidity (NTU)	2.48	2.46	3.10	1.42	2.020000e+00	4.99	2.40
Dissolved Oxygen (mg/L)	6.49	6.49	5.71	2.03	4.110000e+00	7.00	3.54
Nitrate Level (mg/L)	25.08	24.79	14.18	14.51	2.104000e+02	49.94	25.38
Lead Concentration (µg/L)	10.05	10.07	14.70	5.80	3.362000e+01	20.00	9.91
Bacteria Count (CFU/mL)	2488.48	2469.00	4540.00	1431.42	2.048968e+06	4998.00	2468.25
Access to Clean Water (% of Population)	64.61	64.78	44.17	20.31	4.124300e+02	69.98	35.28
Diarrheal Cases per 100,000 people	249.78	248.00	386.00	144.11	2.076814e+04	499.00	254.00
Cholera Cases per 100,000 people	24.25	24.00	3.00	14.33	2.054200e+02	49.00	25.00
Typhoid Cases per 100,000 people	49.27	49.00	56.00	28.98	8.400800e+02	99.00	51.00
Infant Mortality Rate (per 1,000 live births)	50.81	50.23	16.11	28.47	8.102700e+02	97.93	49.79
GDP per Capita (USD)	50036.20	49621.50	5696.00	28598.75	8.178885e+08	99427.00	49768.00
Healthcare Access Index (0-100)	50.03	50.39	0.95	28.90	8.350200e+02	99.79	49.83
Urbanization Rate (%)	50.06	49.80	12.75	22.78	5.188900e+02	79.95	39.17
Sanitation Coverage (% of Population)	60.37	60.58	22.03	23.16	5.363700e+02	79.98	39.98
Rainfall (mm per year)	1591.85	1572.00	283.00	817.50	6.683102e+05	2799.00	1442.50
Temperature (°C)	20.13	20.18	6.86	11.69	1.366400e+02	39.93	20.83
Population Density (people per km²)	505.39	513.00	260.00	283.28	8.024485e+04	989.00	490.25

Prepared by: Rudri Pandya

Water Contamination Levels by Country and Year



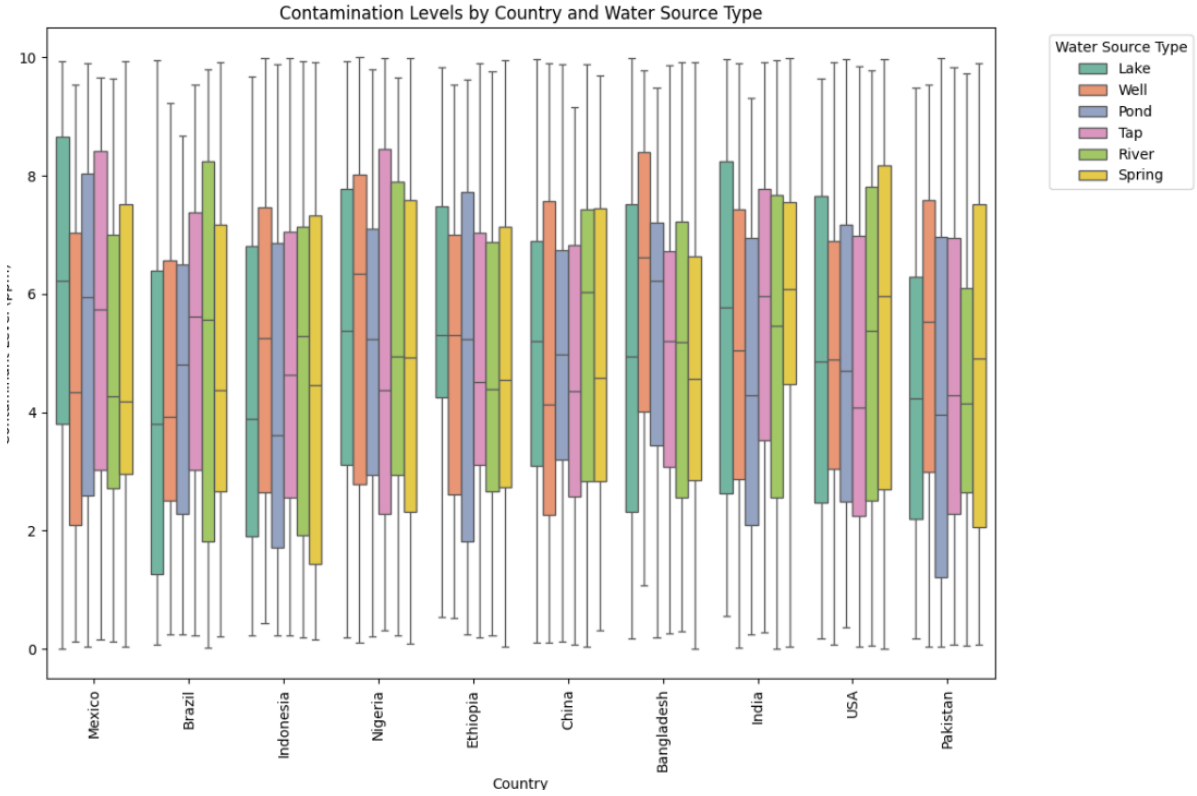
Water Quality Index by Country Over 5-Year Ranges



Prepared by: Rudri Pandya

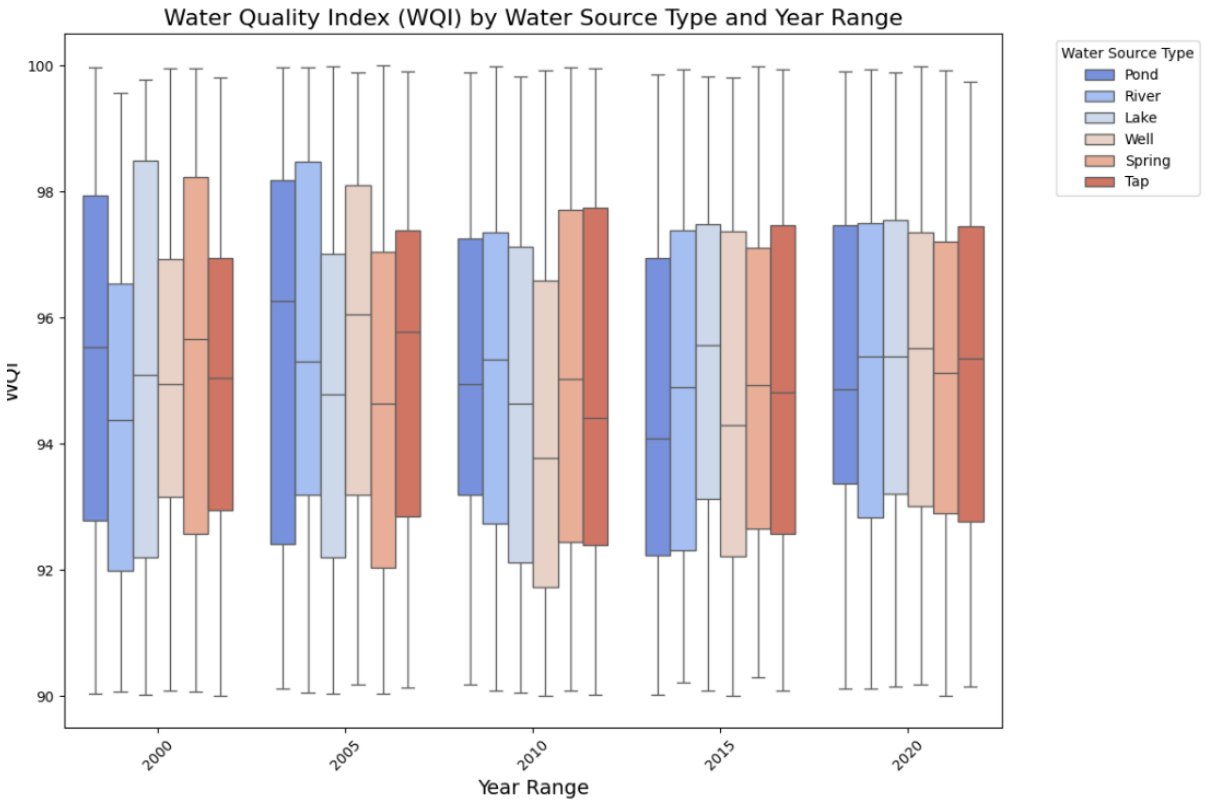
Count of Water Source Types by Country

Contamination Levels by Country and Water Source Type

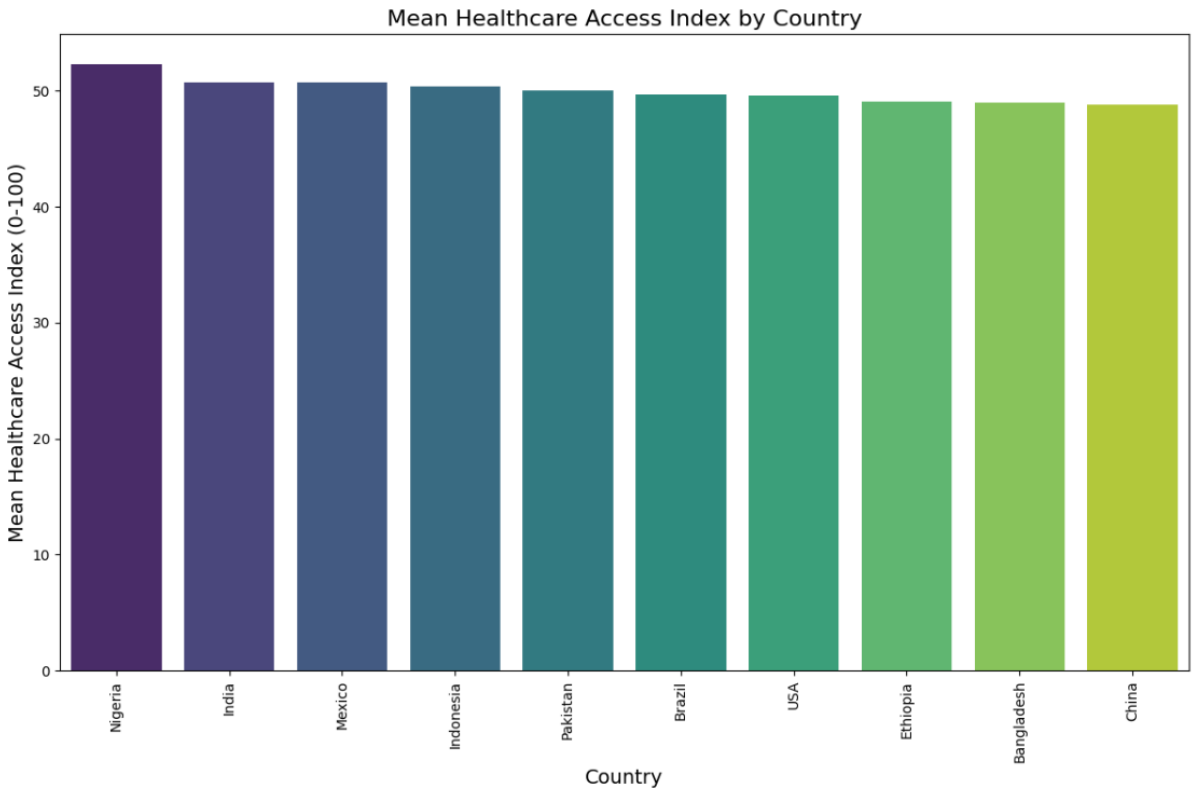


Prepared by: Rudri Pandya

Water Quality Index (WQI) by Water Source Type and Year Range

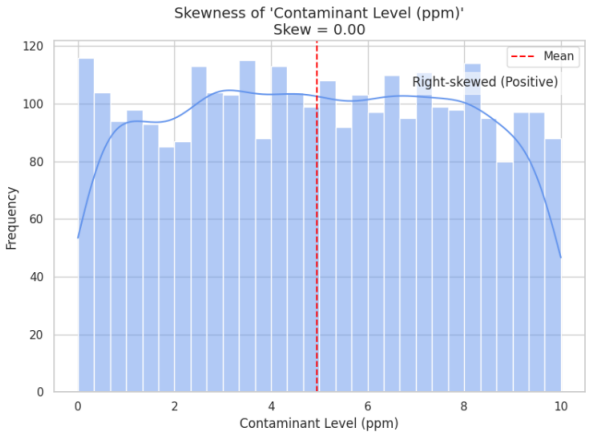


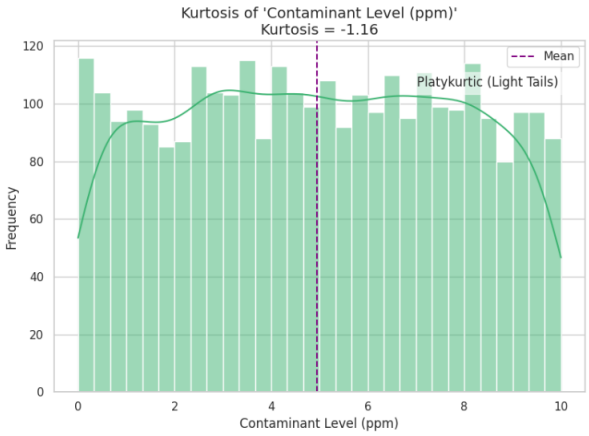
Healthcare Access Index by Country



Prepared by: Rudri Pandya

SKEWNESS AND KURTOSIS





INFERENTIAL STATISTICS- T TEST

- T-Test on Residuals:
- T-statistic: 0.1369
- P-value: 0.8911
- Result: Not Significant — Mean prediction error is not significantly different from zero.

MACHINE LEARNING ALGORITHM -LINEAR REGRESSION

Linear Regression Results:

- Mean Squared Error (MSE): 7.9123
- R^2 Score: -0.0022

T-Test on Residuals:

- T-statistic: 0.1369
- P-value: 0.8911

Result: Not Significant — Mean prediction error is not significantly different from zero.

Coefficients of the Model:

- Turbidity (NTU) -0.009982
- pH Level -0.019396
- Dissolved Oxygen (mg/L) -0.007971
- Nitrate Level (mg/L) -0.010333
- Lead Concentration (µg/L) 0.009726
- Temperature (°C) 0.000433
- Rainfall (mm per year) -0.000010

Conclusion

- Through this comprehensive analysis of water quality and contamination, several key insights emerge:
- Bangladesh, India, and Pakistan face the greatest challenges in water safety, with high contamination levels, low Water Quality Index (WQI), and limited healthcare access. These regions are heavily reliant on surface water and unprotected wells, which contribute to poor water quality and health risks.
- Thailand and China demonstrate strong water infrastructure and healthcare systems, resulting in consistently high WQI scores and better public health outcomes. These countries have effectively implemented piped water systems and improved water treatment, significantly reducing contamination.
- Surface water and unprotected wells are consistently linked to poor water quality across multiple countries. These sources contribute to high contamination levels and are often associated with inadequate sanitation and industrial runoff.
- Piped water systems and rainwater harvesting provide safer alternatives to surface water, especially in rural and developing areas. Countries like Vietnam and Sri Lanka, which utilize rainwater harvesting, demonstrate moderate WQI scores and offer a viable solution where piped water infrastructure is limited.

Overall, this analysis emphasizes the critical need for improved water infrastructure, particularly in developing nations, to ensure safe drinking water and protect public health.

Faculty Name: Dr. Sheetal Pandya

Signature: _____

Prepared by: Rudri Pandya