Analyzing the Impact of Water Pollution on Waterborne Disease Prevalence

This project explores the relationship between water quality indicators and the prevalence of waterborne diseases (such as diarrhea, cholera, and typhoid) across countries and regions from 2000 to 2025. By leveraging the "Water Pollution and Disease" dataset, the analysis aims to identify patterns and correlations between contaminant levels, access to clean water, sanitation coverage, and public health outcomes. The project utilizes statistical analysis and data visualization to uncover key drivers of disease outbreaks and inform potential interventions.

Kaggle Dataset: https://www.kaggle.com/datasets/khushikyad001/water-pollution-and-disease

Github Repo https://github.com/casanchezbar/water_pollution_analysis

Problem Statement

The core problem addressed is understanding how environmental water quality factors contribute to the incidence of waterborne diseases globally.

- Which water quality parameters (e.g., contaminant level, bacteria count, turbidity) are most strongly associated with disease outbreaks?
- How do socioeconomic and infrastructural factors (e.g., sanitation coverage, healthcare access, GDP per capita) modulate these relationships?

Type of problem

- Primarily exploratory data analysis (EDA) and statistical correlation.
- The dataset is also suitable for supervised machine learning tasks, such as regression or classification, to predict disease incidence based on water and socioeconomic features

```python
# load libraries
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import pandas as pd

import statsmodels.api as sm

# Train-Test Split
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import r2_score, mean_squared_error
```

1. Exploratory Data Analysis

```
# Load the data
df =
pd.read_csv('/kaggle/input/water-pollution-and-disease/water_pollution
_disease.csv')
# Display basic information
print('Dataframe info: \n')
display(df.info())

Dataframe info: 

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 24 columns):
 #   Column                                    Non-Null Count
Dtype
---  ------                                    --------------
-----
 0   Country                                   3000 non-null
object
 1   Region                                    3000 non-null
object
 2   Year                                      3000 non-null
int64
 3   Water Source Type                         3000 non-null
object
 4   Contaminant Level (ppm)                   3000 non-null
float64
 5   pH Level                                  3000 non-null
float64
 6   Turbidity (NTU)                           3000 non-null
float64
 7   Dissolved Oxygen (mg/L)                   3000 non-null
float64
 8   Nitrate Level (mg/L)                      3000 non-null
float64
 9   Lead Concentration (µg/L)                 3000 non-null
float64
 10  Bacteria Count (CFU/mL)                   3000 non-null
int64
 11  Water Treatment Method                    2253 non-null
object
 12  Access to Clean Water (% of Population)   3000 non-null
float64
 13  Diarrheal Cases per 100,000 people        3000 non-null
int64
 14  Cholera Cases per 100,000 people          3000 non-null
int64
 15  Typhoid Cases per 100,000 people          3000 non-null
```

```
int64
 16  Infant Mortality Rate (per 1,000 live births)  3000 non-null
float64
 17  GDP per Capita (USD)                           3000 non-null
int64
 18  Healthcare Access Index (0-100)                3000 non-null
float64
 19  Urbanization Rate (%)                          3000 non-null
float64
 20  Sanitation Coverage (% of Population)          3000 non-null
float64
 21  Rainfall (mm per year)                         3000 non-null
int64
 22  Temperature (°C)                               3000 non-null
float64
 23  Population Density (people per km²)            3000 non-null
int64
dtypes: float64(12), int64(8), object(4)
memory usage: 562.6+ KB

None
```

```
# Clean Data

# drop columns not needed
# Water Treatment Method -> high nullability
# Country, Region, Year are irrelavant to our analysis
df = df.drop(['Water Treatment Method','Region','Country','Year'],
axis=1)
df.rename(columns={
    'Water Source Type': 'water_source',
    'Contaminant Level (ppm)': 'contaminant_level',
    'pH Level': 'ph_level',
    'Turbidity (NTU)': 'turbidity',
    'Dissolved Oxygen (mg/L)': 'dissolved_oxygen',
    'Nitrate Level (mg/L)': 'nitrate_level',
    'Lead Concentration (µg/L)': 'lead_concentration',
    'Bacteria Count (CFU/mL)': 'bacteria_count',
    'Access to Clean Water (% of Population)':
'access_to_clean_water',
    'Diarrheal Cases per 100,000 people': 'diarrheal_cases',
    'Cholera Cases per 100,000 people': 'cholera_cases',
    'Typhoid Cases per 100,000 people': 'typhoid_cases',
    'Infant Mortality Rate (per 1,000 live births)':
'infant_mortality_rate',
    'GDP per Capita (USD)': 'gdp',
    'Healthcare Access Index (0-100)': 'healthcare_access',
    'Urbanization Rate (%)': 'urbanization_rate',
    'Sanitation Coverage (% of Population)': 'sanitation_coverage',
    'Rainfall (mm per year)': 'rainfall_per_year',
```

```
    'Temperature (°C)': 'temperature',
    'Population Density (people per km²)': 'population_density'
}, inplace=True)

df = pd.get_dummies(df, drop_first=True)

bool_cols = ['water_source_Pond', 'water_source_River',
'water_source_Spring', 'water_source_Tap', 'water_source_Well']
df[bool_cols] = df[bool_cols].astype(int)

df

      Year  contaminant_level  ph_level  turbidity
dissolved_oxygen  \
0     2015               6.06      7.12       3.93                 4.28

1     2017               5.24      7.84       4.79                 3.86

4     2005               0.12      8.16       4.22                 9.15

7     2024               3.76      6.42       1.35                 9.99

10    2023               4.16      8.43       4.61                 6.25

...    ...                ...       ...        ...                  ...

2967  2013               0.54      6.07       4.94                 3.14

2979  2001               9.51      8.30       0.76                 6.66

2981  2005               0.28      6.50       3.40                 3.52

2993  2009               1.98      7.71       1.26                 6.30

2999  2013               0.98      7.69       2.55                 7.42


      nitrate_level  lead_concentration  bacteria_count  \
0              8.28                7.89            3344
1             15.74               14.68            2122
4             49.35               12.51            4182
7              2.73               19.44            1172
10            14.36               11.20            2129
...             ...                 ...             ...
2967           0.54                8.51            3749
2979          40.00                0.85            4194
2981          47.75               13.19            1078
2993          41.30               13.18            1171
2999          19.78                1.04            1719

      access_to_clean_water  diarrheal_cases  ...  sanitation_coverage
```

```
\
0                           33.60              472  ...                  63.23

1                           89.54              122  ...                  29.12

4                           36.60              466  ...                  69.23

7                           44.17              397  ...                  70.22

10                          41.68              286  ...                  25.37

...                           ...              ...  ...                    ...

2967                        92.09              200  ...                  34.48

2979                        53.43              123  ...                  36.61

2981                        46.08              197  ...                  57.93

2993                        38.12               55  ...                  84.55

2999                        49.37              440  ...                  67.31


      rainfall_per_year  temperature  population_density
Country_Mexico  \
0                  2800         4.94                 593
True
1                  1572        16.93                 234
False
4                  2295        31.44                 414
True
7                   940        19.64                 111
True
10                 1144         9.54                 299
False
...                 ...          ...                 ...              .
..
2967                774        23.56                 134
True
2979               2661         0.29                 848
True
2981               2055        37.68                 598
False
2993                718        36.89                 719
True
2999                937         9.60                 274
True

      water_source_Pond  water_source_River  water_source_Spring  \
```

```
0                      0                    0                     0
1                      0                    0                     0
4                      0                    0                     0
7                      0                    0                     1
10                     0                    0                     0
...                  ...                  ...                   ...
2967                   0                    0                     0
2979                   0                    0                     0
2981                   0                    0                     0
2993                   0                    0                     0
2999                   0                    0                     1

      water_source_Tap  water_source_Well
0                    0                  0
1                    0                  1
4                    0                  1
7                    0                  0
10                   0                  1
...                ...                ...
2967                 1                  0
2979                 0                  1
2981                 0                  0
2993                 0                  1
2999                 0                  0

[579 rows x 26 columns]
```

Summary Statistics

```
print('Dataframe num summary statistics: \n')
display(df.describe())

# Display the first few rows
df.head()
```

```
Dataframe num summary statistics:


       contaminant_level     ph_level    turbidity
dissolved_oxygen  \
count        3000.000000  3000.000000  3000.000000       3000.000000

mean            4.954390     7.255847     2.480023          6.492850

std             2.860072     0.720464     1.419984          2.027966

min             0.000000     6.000000     0.000000          3.000000

25%             2.560000     6.630000     1.257500          4.710000
```

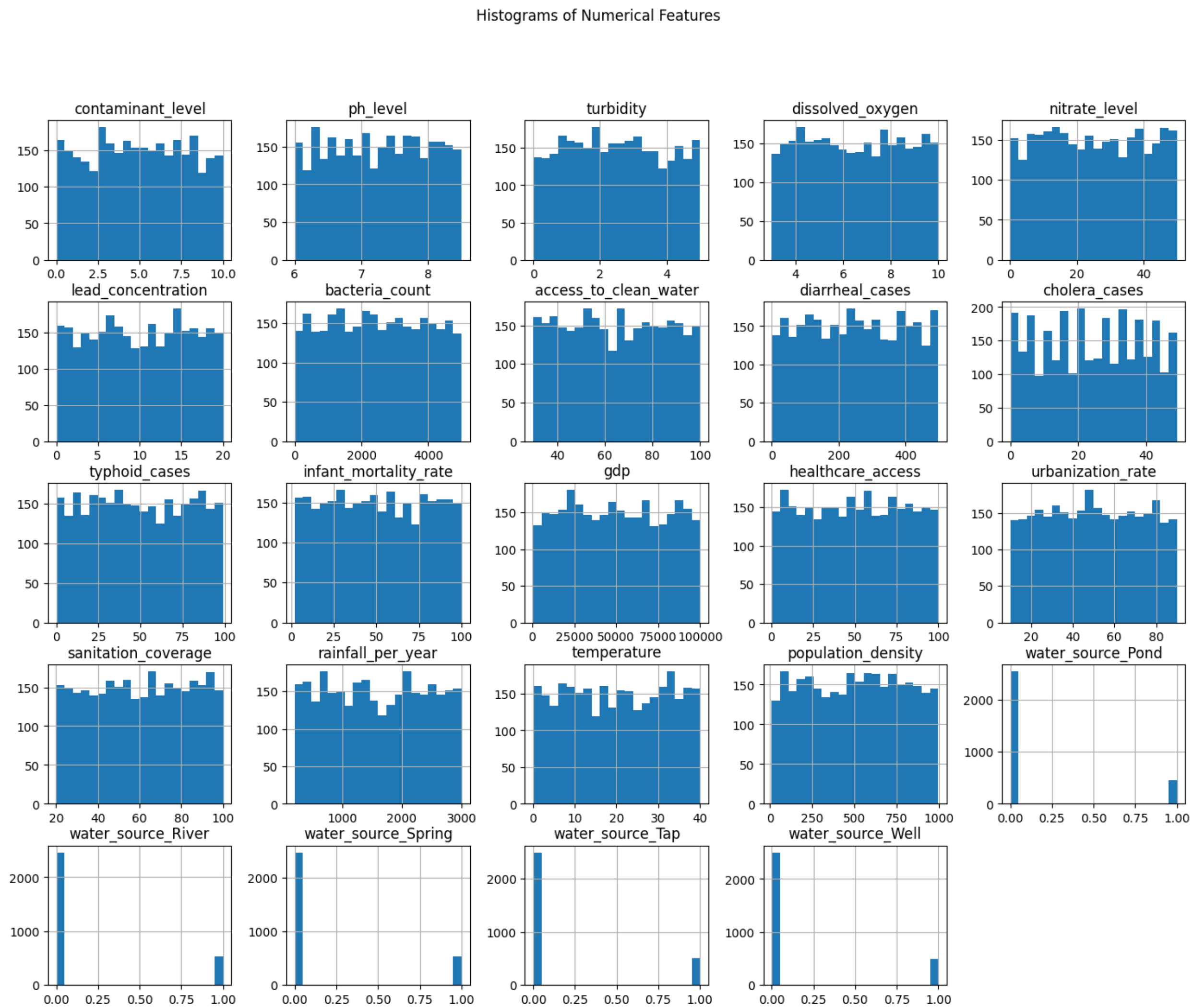
```
50%            4.950000     7.280000     2.460000         6.490000

75%            7.400000     7.870000     3.660000         8.252500

max           10.000000     8.500000     4.990000        10.000000


       nitrate_level  lead_concentration  bacteria_count  \
count     3000.00000         3000.000000     3000.000000
mean        25.08025           10.047913     2488.477333
std         14.50517            5.798238     1431.421553
min          0.05000            0.000000        0.000000
25%         12.52500            5.120000     1268.000000
50%         24.79000           10.065000     2469.000000
75%         37.91000           15.032500     3736.250000
max         49.99000           20.000000     4998.000000

       access_to_clean_water  diarrheal_cases  cholera_cases  ...  \
count            3000.000000      3000.000000     3000.00000  ...
mean               64.612333       249.776667       24.25100  ...
std                20.308463       144.111543       14.33259  ...
min                30.010000         0.000000        0.00000  ...
25%                47.027500       124.000000       12.00000  ...
50%                64.780000       248.000000       24.00000  ...
75%                82.302500       378.000000       37.00000  ...
max                99.990000       499.000000       49.00000  ...

       urbanization_rate  sanitation_coverage  rainfall_per_year
temperature  \
count        3000.000000          3000.000000        3000.000000
3000.000000
mean           50.062480            60.371007        1591.849000
20.130917
std            22.779125            23.159678         817.502434
11.689244
min            10.030000            20.010000         200.000000
0.060000
25%            30.557500            40.440000         865.750000
9.840000
50%            49.795000            60.580000        1572.000000
20.175000
75%            69.727500            80.420000        2308.250000
30.672500
max            89.980000            99.990000        2999.000000
39.990000

       population_density  water_source_Pond  water_source_River  \
count         3000.000000        3000.000000         3000.000000
mean           505.390333           0.150000            0.179333
```

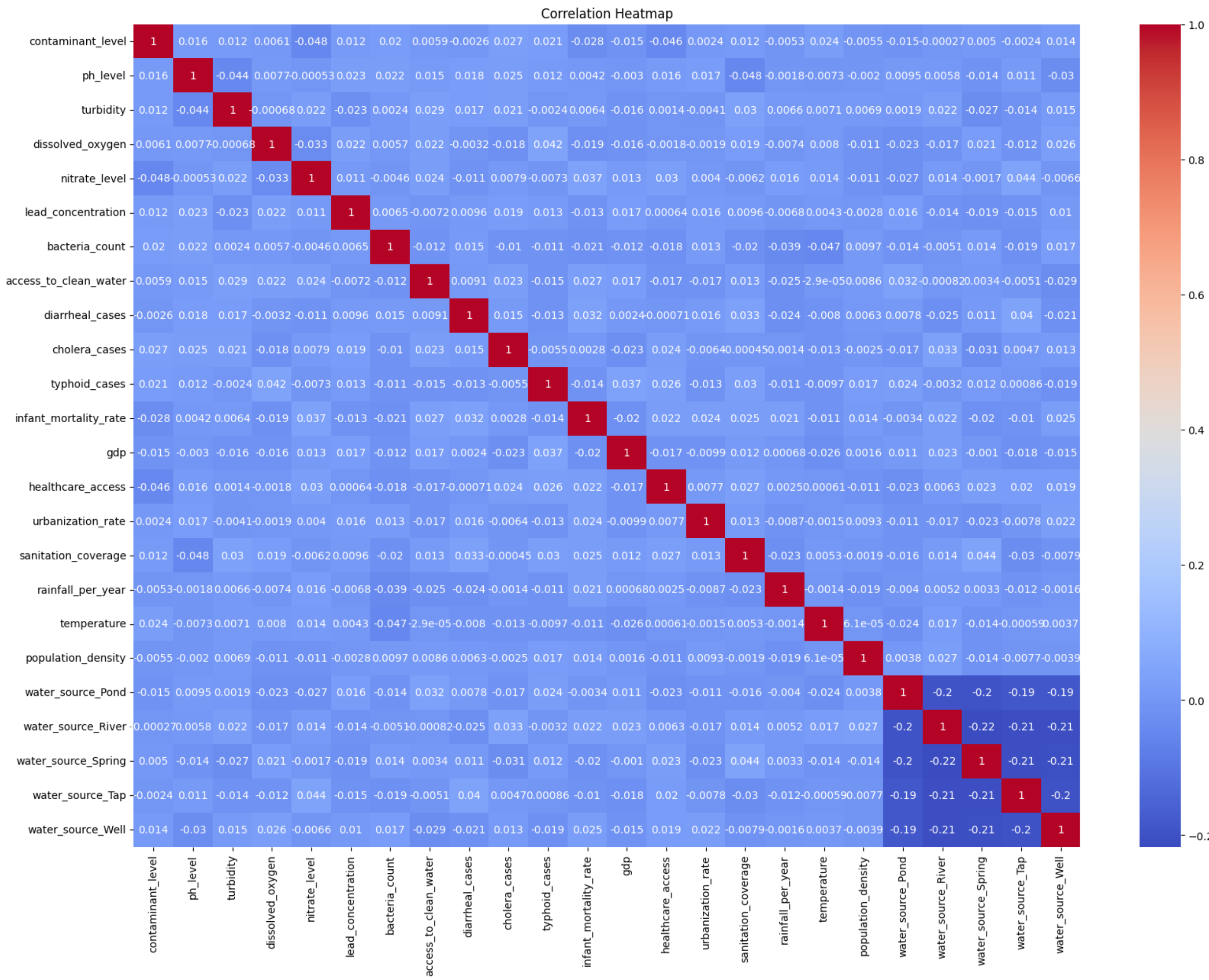
```
std         283.275224         0.357131          0.383695
min          10.000000         0.000000          0.000000
25%         254.750000         0.000000          0.000000
50%         513.000000         0.000000          0.000000
75%         745.000000         0.000000          0.000000
max         999.000000         1.000000          1.000000

       water_source_Spring  water_source_Tap  water_source_Well
count          3000.000000       3000.000000        3000.000000
mean              0.177333          0.167000           0.166000
std               0.382014          0.373038           0.372143
min               0.000000          0.000000           0.000000
25%               0.000000          0.000000           0.000000
50%               0.000000          0.000000           0.000000
75%               0.000000          0.000000           0.000000
max               1.000000          1.000000           1.000000

[8 rows x 24 columns]

   contaminant_level  ph_level  turbidity  dissolved_oxygen
nitrate_level  \
0               6.06      7.12       3.93              4.28
8.28
1               5.24      7.84       4.79              3.86
15.74
2               0.24      6.43       0.79              3.42
36.67
3               7.91      6.71       1.96              3.12
36.92
4               0.12      8.16       4.22              9.15
49.35

   lead_concentration  bacteria_count  access_to_clean_water
diarrheal_cases  \
0                7.89            3344                  33.60
472
1               14.68            2122                  89.54
122
2                9.96            2330                  35.29
274
3                6.77            3779                  57.53
3
4               12.51            4182                  36.60
466

   cholera_cases  ...  urbanization_rate  sanitation_coverage  \
0             33  ...              84.61                63.23
1             27  ...              73.37                29.12
2             39  ...              72.86                93.56
3             33  ...              71.07                94.25
```

```
4              31  ...              55.55                69.23

   rainfall_per_year  temperature  population_density
water_source_Pond  \
0               2800         4.94                 593
0
1               1572        16.93                 234
0
2               2074        21.73                  57
1
3                937         3.79                 555
0
4               2295        31.44                 414
0

   water_source_River  water_source_Spring  water_source_Tap  \
0                   0                    0                 0
1                   0                    0                 0
2                   0                    0                 0
3                   0                    0                 0
4                   0                    0                 0

   water_source_Well
0                  0
1                  1
2                  0
3                  1
4                  1

[5 rows x 24 columns]
```

Data Visualization

```
# Histograms
df.hist(figsize=(17, 13), bins=20)
plt.suptitle("Histograms of Numerical Features")
plt.show()

# Correlation heatmap
plt.figure(figsize=(20, 14))
sns.heatmap(df.corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.show()
```


Histograms of Numerical Features
contaminant_level
ph_level
turbidity
dissolved_oxygen
nitrate_level
lead_concentration
bacteria_count
access_to_clean_water
diarrheal_cases
cholera_cases
typhoid_cases
infant_mortality_rate
gdp
healthcare_access
urbanization_rate
sanitation_coverage
rainfall_per_year
temperature
population_density
water_source_Pond
water_source_River
water_source_Spring
water_source_Tap
water_source_Well



Modeling

Water quality

- **Selected features for water quality analysis**:
  - 'contaminant_level',
  - 'ph_level',
  - 'turbidity',
  - 'dissolved_oxygen',
  - 'nitrate_level',
  - 'lead_concentration',
  - 'bacteria_count',
  - 'access_to_clean_water',
  - 'diarrheal_cases',
  - 'cholera_cases',
  - 'typhoid_cases',

Socioeconomic

- **Selected features for socioeconomic analysis**:
  - 'water_source',
  - 'access_to_clean_water'
  - 'infant_mortality_rate',
  - 'gdp',
  - 'healthcare_access',
  - 'urbanization_rate',
  - 'sanitation_coverage',
  - 'rainfall_per_year',
  - 'temperature'

```
water_quality_features =
['contaminant_level','ph_level','turbidity','dissolved_oxygen','nitrat
e_level', 'lead_concentration', 'bacteria_count', 'diarrheal_cases',
'cholera_cases','typhoid_cases',]
df_water_quality = df[water_quality_features]

socio_features =
['water_source_Pond','water_source_River','water_source_Spring','water
_source_Tap','water_source_Well','access_to_clean_water','infant_morta
lity_rate','gdp','healthcare_access', 'urbanization_rate',
'sanitation_coverage','rainfall_per_year',
'temperature','diarrheal_cases','population_density',
'cholera_cases','typhoid_cases',]
df_socio = df[socio_features]

for decease in ['diarrheal','cholera','typhoid']:
    for current_df in [df_water_quality,df_socio]:
        # Define features and target for diarrheal
        X =
current_df.drop(["diarrheal_cases",'cholera_cases','typhoid_cases'],
axis=1)
        y = current_df[f"{decease}_cases"]
        X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

        model = sm.OLS(y_train, X_train)
        results = model.fit()
        print(results.summary())

        # Predict on test data
        y_pred = results.predict(X_test)

        # Evaluate on test data
        r2 = r2_score(y_test, y_pred)
        rmse = mean_squared_error(y_test, y_pred, squared=False)
```

```
    print(f"Test R^2: {r2:.3f}")
    print(f"Test RMSE: {rmse:.3f}")
```

```
                                 OLS Regression Results

==============================================================================
=================
Dep. Variable:          diarrheal_cases   R-squared (uncentered):
0.747
Model:                              OLS   Adj. R-squared (uncentered):
0.746
Method:                   Least Squares   F-statistic:
1009.
Date:                  Fri, 02 May 2025   Prob (F-statistic):
0.00
Time:                          04:17:04   Log-Likelihood:
-15345.
No. Observations:                  2400   AIC:
3.070e+04
Df Residuals:                      2393   BIC:
3.075e+04
Df Model:                             7

Covariance Type:              nonrobust

==============================================================================
================
                         coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
----------------
contaminant_level     -0.1652      1.019     -0.162      0.871      -
2.163       1.833
ph_level              27.5570      1.885     14.622      0.000
23.861      31.253
turbidity              5.6979      2.045      2.787      0.005
1.688       9.708
dissolved_oxygen       2.4526      1.407      1.743      0.081      -
0.306       5.211
nitrate_level          0.2085      0.203      1.029      0.304      -
0.189       0.606
lead_concentration     0.6561      0.508      1.291      0.197      -
0.340       1.652
bacteria_count         0.0026      0.002      1.247      0.213      -
0.001       0.007
==============================================================================
========
Omnibus:                       1004.843   Durbin-Watson:
2.015
Prob(Omnibus):                    0.000   Jarque-Bera (JB):
```

```
123.478
Skew:                           0.032   Prob(JB):
1.54e-27
Kurtosis:                       1.891   Cond. No.
2.25e+03
======================================================================
========

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[3] The condition number is large, 2.25e+03. This might indicate that
there are
strong multicollinearity or other numerical problems.
Test R^2: -0.027
Test RMSE: 147.720
                                 OLS Regression Results

======================================================================
=================
Dep. Variable:        diarrheal_cases   R-squared (uncentered):
0.739
Model:                            OLS   Adj. R-squared (uncentered):
0.738
Method:                 Least Squares   F-statistic:
483.5
Date:                Fri, 02 May 2025   Prob (F-statistic):
0.00
Time:                        04:17:04   Log-Likelihood:
-15381.
No. Observations:                2400   AIC:
3.079e+04
Df Residuals:                    2386   BIC:
3.087e+04
Df Model:                          14

Covariance Type:            nonrobust

======================================================================
===================
                            coef    std err          t      P>|t|
[0.025      0.975]
----------------------------------------------------------------------
-------------------
water_source_Pond        35.3924     10.619      3.333      0.001
14.569      56.216
water_source_River       28.4013     10.096      2.813      0.005
8.603     48.199
```

```
water_source_Spring      29.4941     10.087      2.924      0.003
9.714      49.275
water_source_Tap         49.0714     10.191      4.815      0.000
29.087      69.055
water_source_Well        19.6676     10.167      1.935      0.053
-0.269      39.604
access_to_clean_water     0.7834      0.132      5.946      0.000
0.525       1.042
infant_mortality_rate     0.3665      0.103      3.554      0.000
0.164       0.569
gdp                       0.0003      0.000      2.884      0.004
9.37e-05       0.000
healthcare_access         0.2484      0.101      2.453      0.014
0.050       0.447
urbanization_rate         0.5824      0.125      4.672      0.000
0.338       0.827
sanitation_coverage       0.6660      0.121      5.505      0.000
0.429       0.903
rainfall_per_year         0.0133      0.004      3.757      0.000
0.006       0.020
temperature               0.5776      0.249      2.317      0.021
0.089       1.067
population_density        0.0374      0.010      3.625      0.000
0.017       0.058
======================================================================
========
Omnibus:                      521.031   Durbin-Watson:
2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
99.250
Skew:                           0.024   Prob(JB):
2.81e-22
Kurtosis:                       2.005   Cond. No.
3.37e+05
======================================================================
========

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[3] The condition number is large, 3.37e+05. This might indicate that
there are
strong multicollinearity or other numerical problems.
Test R^2: -0.058
Test RMSE: 149.971
                                 OLS Regression Results

======================================================================
```

```
=================
Dep. Variable:          cholera_cases   R-squared (uncentered):
0.738
Model:                            OLS   Adj. R-squared (uncentered):
0.737
Method:                 Least Squares   F-statistic:
962.3
Date:                Fri, 02 May 2025   Prob (F-statistic):
0.00
Time:                        04:17:04   Log-Likelihood:
-9812.5
No. Observations:                2400   AIC:
1.964e+04
Df Residuals:                    2393   BIC:
1.968e+04
Df Model:                           7

Covariance Type:            nonrobust

======================================================================
================
                         coef    std err          t      P>|t|
[0.025      0.975]
----------------------------------------------------------------------
----------------
contaminant_level      0.2012      0.102      1.980      0.048
0.002       0.400
ph_level               2.5749      0.188     13.701      0.000
2.206       2.943
turbidity              0.4944      0.204      2.424      0.015
0.094       0.894
dissolved_oxygen       0.1445      0.140      1.030      0.303      -
0.131       0.420
nitrate_level          0.0337      0.020      1.668      0.095      -
0.006       0.073
lead_concentration     0.0980      0.051      1.934      0.053      -
0.001       0.197
bacteria_count         0.0002      0.000      0.817      0.414      -
0.000       0.001
======================================================================
========
Omnibus:                     1082.985   Durbin-Watson:
1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
125.658
Skew:                          -0.005   Prob(JB):
5.17e-28
Kurtosis:                       1.879   Cond. No.
2.25e+03
======================================================================
```

```
========

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[3] The condition number is large, 2.25e+03. This might indicate that
there are
strong multicollinearity or other numerical problems.
Test R^2: -0.002
Test RMSE: 14.366
                                 OLS Regression Results

======================================================================
=================
Dep. Variable:          cholera_cases   R-squared (uncentered):
0.728
Model:                            OLS   Adj. R-squared (uncentered):
0.727
Method:                 Least Squares   F-statistic:
457.0
Date:                Fri, 02 May 2025   Prob (F-statistic):
0.00
Time:                        04:17:04   Log-Likelihood:
-9855.3
No. Observations:                2400   AIC:
1.974e+04
Df Residuals:                    2386   BIC:
1.982e+04
Df Model:                          14

Covariance Type:            nonrobust

======================================================================
===================
                            coef    std err          t      P>|t|
[0.025      0.975]
----------------------------------------------------------------------
-------------------
water_source_Pond         3.1934      1.062      3.006      0.003
1.110       5.277
water_source_River        4.0317      1.010      3.992      0.000
2.051       6.012
water_source_Spring       2.1240      1.009      2.105      0.035
0.145       4.103
water_source_Tap          3.7507      1.020      3.679      0.000
1.751       5.750
water_source_Well         3.9481      1.017      3.882      0.000
1.954       5.943
```

```
access_to_clean_water     0.0932      0.013      7.068      0.000
0.067       0.119
infant_mortality_rate     0.0221      0.010      2.137      0.033
0.002       0.042
gdp                    2.517e-05   1.02e-05      2.479      0.013
5.26e-06    4.51e-05
healthcare_access         0.0416      0.010      4.108      0.000
0.022       0.061
urbanization_rate         0.0439      0.012      3.524      0.000
0.019       0.068
sanitation_coverage       0.0571      0.012      4.715      0.000
0.033       0.081
rainfall_per_year         0.0012      0.000      3.536      0.000
0.001       0.002
temperature               0.0556      0.025      2.228      0.026
0.007       0.104
population_density        0.0033      0.001      3.190      0.001
0.001       0.005
======================================================================
========
Omnibus:                      660.647   Durbin-Watson:
1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
107.998
Skew:                          -0.012   Prob(JB):
3.54e-24
Kurtosis:                       1.961   Cond. No.
3.37e+05
======================================================================
========

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[3] The condition number is large, 3.37e+05. This might indicate that
there are
strong multicollinearity or other numerical problems.
Test R^2: -0.067
Test RMSE: 14.822
                                 OLS Regression Results

======================================================================
=================
Dep. Variable:          typhoid_cases   R-squared (uncentered):
0.739
Model:                            OLS   Adj. R-squared (uncentered):
0.738
Method:                 Least Squares   F-statistic:
```

```
968.5
Date:                  Fri, 02 May 2025   Prob (F-statistic):
0.00
Time:                          04:17:04   Log-Likelihood:
-11499.
No. Observations:                  2400   AIC:
2.301e+04
Df Residuals:                      2393   BIC:
2.305e+04
Df Model:                             7

Covariance Type:              nonrobust

======================================================================
================
                         coef    std err          t      P>|t|
[0.025      0.975]
----------------------------------------------------------------------
----------------
contaminant_level      0.2060      0.205      1.004      0.315      -
0.196      0.608
ph_level               5.1371      0.380     13.536      0.000
4.393      5.881
turbidity              0.4748      0.412      1.153      0.249      -
0.333      1.282
dissolved_oxygen       1.2438      0.283      4.391      0.000
0.688      1.799
nitrate_level          0.0281      0.041      0.688      0.491      -
0.052      0.108
lead_concentration     0.0480      0.102      0.470      0.639      -
0.153      0.249
bacteria_count      2.936e-05      0.000      0.071      0.943      -
0.001      0.001
======================================================================
========
Omnibus:                       1243.254   Durbin-Watson:
1.985
Prob(Omnibus):                    0.000   Jarque-Bera (JB):
130.340
Skew:                             0.015   Prob(JB):
4.98e-29
Kurtosis:                         1.859   Cond. No.
2.25e+03
======================================================================
========

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
```

```
correctly specified.
[3] The condition number is large, 2.25e+03. This might indicate that
there are
strong multicollinearity or other numerical problems.
Test R^2: -0.026
Test RMSE: 29.351
                                 OLS Regression Results

======================================================================
=================
Dep. Variable:          typhoid_cases   R-squared (uncentered):
0.729
Model:                            OLS   Adj. R-squared (uncentered):
0.728
Method:                 Least Squares   F-statistic:
459.4
Date:                Fri, 02 May 2025   Prob (F-statistic):
0.00
Time:                        04:17:04   Log-Likelihood:
-11543.
No. Observations:                2400   AIC:
2.311e+04
Df Residuals:                    2386   BIC:
2.319e+04
Df Model:                          14

Covariance Type:            nonrobust

======================================================================
===================
                            coef    std err          t      P>|t|
[0.025      0.975]
----------------------------------------------------------------------
-------------------
water_source_Pond         8.7126      2.146      4.060      0.000
4.504      12.921
water_source_River        6.1582      2.040      3.018      0.003
2.157      10.159
water_source_Spring       7.1858      2.039      3.525      0.000
3.188      11.183
water_source_Tap          6.0250      2.060      2.925      0.003
1.986      10.064
water_source_Well         4.2238      2.055      2.056      0.040
0.195       8.253
access_to_clean_water     0.1282      0.027      4.816      0.000
0.076       0.180
infant_mortality_rate     0.0435      0.021      2.089      0.037
0.003       0.084
gdp                       0.0001   2.05e-05      5.116      0.000
6.47e-05       0.000
```

```
healthcare_access         0.0791      0.020      3.864      0.000
0.039      0.119
urbanization_rate         0.0837      0.025      3.324      0.001
0.034      0.133
sanitation_coverage       0.1536      0.024      6.281      0.000
0.106      0.202
rainfall_per_year         0.0017      0.001      2.323      0.020
0.000      0.003
temperature               0.1347      0.050      2.674      0.008
0.036      0.234
population_density        0.0083      0.002      3.977      0.000
0.004      0.012
======================================================================
========
Omnibus:                    577.423   Durbin-Watson:
2.010
Prob(Omnibus):                0.000   Jarque-Bera (JB):
103.034
Skew:                         0.019   Prob(JB):
4.23e-23
Kurtosis:                     1.986   Cond. No.
3.37e+05
======================================================================
========

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[3] The condition number is large, 3.37e+05. This might indicate that
there are
strong multicollinearity or other numerical problems.
Test R^2: -0.055
Test RMSE: 29.757
```

Water quality analysis

Diarrheal Cases Model result

- Strong model fit with an uncentered R-squared of 0.747, indicating that approximately 74.7% of the variation in diarrheal_cases is explained by the model without including an intercept
- **ph_level**: Highly significant ($p<0.001$). Strong positive association: as pH increases by 1 unit, diarrheal cases increase by ~27.56 units, holding other factors constant.
- **turbidity**: Statistically significant ($p=0.005$). Positive effect: higher turbidity increases diarrheal cases.
- **dissolved_oxygen**: Marginally insignificant ($p=0.081$). Suggestive positive trend but not conclusive at 0.05 level.

Cholera Cases Model result

- Model explains ~73.8% of the variance in cholera cases (without an intercept term)
- **contaminant_level**: Statistically significant ($p<0.05$); For each unit increase in contaminant level, cholera cases increase by 0.2012
- **ph_level**: Highly significant ($p<0.001$); Each unit increase in pH, cholera cases increase by ~2.5749
- **turbidity**: Statistically significant ($p<0.05$); Each unit increase in turbidity, cholera cases increase by 0.4944
- **lead_concentration**: Marginally significant (p slightly > 0.05); may warrant further exploration

Typhoid Cases Model result

- Model explains ~73.9% of variance (no intercept)
- **ph_level**: Highly significant ($p<0.001$). Strong positive effect on typhoid cases.
- **dissolved_oxygen**: Statistically significant ($p<0.001$). Positive effect on typhoid cases.

Socioeconomical Analysis

Diarrheal Cases Model result

- Strong model fit with an uncentered R-squared of 0.739, indicating that approximately 73.9% of the variation in diarrheal_cases is explained by the model without including an intercept
- **water_source_Pond**: Significant, using pond water associated with ~35 more diarrheal cases compared to baseline.
- **water_source_River**: Significant; river water use linked to increased diarrheal cases.
- **water_source_Spring**: Significant; spring water use also associated with higher cases.
- **water_source_Tap**: Highly significant; tap water use linked to even higher diarrheal cases (possibly a data artifact or contamination issue).
- **water_source_Well**:Marginally significant; well water may be associated with increased cases.
- **access_to_clean_water**: Significant positive coefficient; counterintuitive and may indicate confounding or coding issues.
- **infant_mortality_rate**: Significant; higher infant mortality correlates with more diarrheal cases.
- **gdp**:Significant but very small effect size.
- **healthcare_access**: Significant; better healthcare access linked with increased reported cases (could reflect reporting bias).
- **urbanization_rate**: Significant; urbanization associated with higher diarrheal cases.
- **sanitation_coverage**: Significant; higher sanitation coverage surprisingly linked to more cases (may indicate confounding).
- **rainfall_per_year**: Significant; more rainfall associated with increased cases.
- **temperature**: Significant; higher temperature linked to more diarrheal cases.
- **population_density**: Significant; denser populations have more cases.

Cholera Cases Model result

- Model explains ~72.8% of the variance in cholera cases (without an intercept term)
- All predictors have positive coefficients and are statistically significant, indicating consistent positive associations with the outcome.
- The water source variables show moderate increases in the outcome, suggesting that the type of water source significantly impacts the dependent variable.
- Socioeconomic and environmental factors like access to clean water, infant mortality, healthcare access, urbanization, sanitation, rainfall, temperature, and population density all have positive but varying effect sizes.
- The effect sizes for some variables (e.g., GDP, rainfall, population density) are small but statistically significant, likely reflecting subtle influences.cases.

Typhoid Cases Model

- All predictors are statistically significant at the 5% level.
- Water source types have substantial positive associations with typhoid cases.
- Socioeconomic and environmental variables (access to clean water, infant mortality, healthcare access, urbanization, sanitation, rainfall, temperature, population density) all show positive effects.
- Positive coefficients for access to clean water and sanitation coverage might indicate:
- Reporting bias (better infrastructure areas report more cases)
- Confounding factors

Conclusions

Water Quality Analysis

Diarrheal Cases Model

- The model demonstrates a strong fit with an uncentered R-squared of 0.747, explaining approximately 74.7% of the variation in diarrheal cases without including an intercept.
- pH level is highly significant ($p < 0.001$) with a strong positive association: a 1-unit increase in pH corresponds to an increase of about 27.56 diarrheal cases, holding other factors constant.
- Turbidity is also statistically significant ($p = 0.005$), indicating that higher turbidity levels increase diarrheal cases.
- Dissolved oxygen shows a suggestive positive trend ($p = 0.081$) but is marginally insignificant at the 0.05 level.

Cholera Cases Model

- The model explains about 73.8% of the variance in cholera cases (uncentered R-squared).
- Contaminant level is statistically significant ($p < 0.05$), with each unit increase associated with a 0.2012 increase in cholera cases.
- pH level remains highly significant ($p < 0.001$), with a positive effect size of approximately 2.57 cases per unit increase.
- Turbidity is significant ($p < 0.05$), positively associated with cholera cases.

- Lead concentration is marginally significant (p slightly > 0.05), suggesting further investigation may be warranted.

Typhoid Cases Model

- The model explains approximately 73.9% of the variance (uncentered R-squared).
- Both pH level and dissolved oxygen are highly significant ($p < 0.001$), with positive effects on typhoid cases.

Socioeconomic Analysis

Diarrheal Cases Model

- The model shows a strong fit with an uncentered R-squared of 0.739.
- Various water sources (pond, river, spring, tap, well) are significant predictors, with pond and tap water associated with notably higher diarrheal cases.
- Access to clean water has a significant positive coefficient, which is counterintuitive and may indicate confounding or data coding issues.
- Other significant socioeconomic factors include:
  - Infant mortality rate
  - GDP (small effect size)
  - Healthcare access (positive association possibly due to reporting bias)
  - Urbanization rate
  - Sanitation coverage (positive association, potentially confounded)
  - Rainfall per year
  - Temperature
  - Population density

Cholera Cases Model

- The model explains approximately 72.8% of the variance.
- All predictors have positive and statistically significant coefficients.
- Water source types moderately increase cholera cases, highlighting the impact of water quality.
- Socioeconomic and environmental factors (access to clean water, infant mortality, healthcare access, urbanization, sanitation, rainfall, temperature, population density) also positively influence cholera cases, though some effect sizes (e.g., GDP, rainfall) are small.

Typhoid Cases Model

- All predictors are statistically significant at the 5% level.
- Water source types show substantial positive associations with typhoid cases.
- Socioeconomic and environmental variables similarly exhibit positive effects.
- Positive coefficients for access to clean water and sanitation coverage may reflect:
- Reporting bias (areas with better infrastructure report more cases)
- Confounding factors requiring further study.

- Overall Insights: Water quality indicators such as pH, turbidity, contaminant levels, and dissolved oxygen play crucial roles in explaining the incidence of waterborne diseases.
- Socioeconomic and environmental factors significantly contribute to disease prevalence, but some counterintuitive findings (e.g., positive effects of clean water access and sanitation coverage) suggest the influence of confounding variables or reporting biases.

The models demonstrate strong explanatory power but highlight the need for careful interpretation and further investigation into data quality, confounding, and potential causal mechanisms.