In [1]:
```python
# Import required libraries
import os
import pandas as pd
import numpy as np
import logging
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np


# Define file paths
file_paths = [
    '/Users/a/Documents/DataScience_World/Regonet/last_project//Infant_Mortality_Dataset/1_youth-mortality-rate
    '/Users/a/Documents/DataScience_World/Regonet/last_project//Infant_Mortality_Dataset/2_number-of-infant-dea
    '/Users/a/Documents/DataScience_World/Regonet/last_project//Infant_Mortality_Dataset/3_child-mortality-by-i
    '/Users/a/Documents/DataScience_World/Regonet/last_project//Infant_Mortality_Dataset/4_Distribution of Caus
    '/Users/a/Documents/DataScience_World/Regonet/last_project//Infant_Mortality_Dataset/5_number-of-maternal-d
    '/Users/a/Documents/DataScience_World/Regonet/last_project//Infant_Mortality_Dataset/6_births-attended-by-h
    '/Users/a/Documents/DataScience_World/Regonet/last_project//Infant_Mortality_Dataset/7_global-vaccination-c
    '/Users/a/Documents/DataScience_World/Regonet/last_project//Infant_Mortality_Dataset/8_health-protection-co
]

# Validate file paths
def validate_paths(file_paths):
    """
    Validate that all file paths exist

    Args:
        file_paths (list): List of file paths

    Returns:
        bool: True if all paths are valid, False otherwise
    """
    valid_paths = all(os.path.exists(path) for path in file_paths)
    if not valid_paths:
        print("Error: One or more file paths are invalid!")
    return valid_paths

if not validate_paths(file_paths):
    raise FileNotFoundError("Invalid file paths! Please check the paths and try again.")
```

In [2]:
```python
# Load datasets
def load_datasets(file_paths):
    """
    Load datasets from file paths

    Args:
        file_paths (list): List of file paths

    Returns:
        dict: Dictionary of loaded datasets
    """
    datasets = {}
    for i, path in enumerate(file_paths, 1):
        try:
            df = pd.read_csv(path, low_memory=False)
            dataset_name = f"data_{i}_{os.path.basename(path).split('.')[0]}"
            datasets[dataset_name] = df
            print(f"Loaded {dataset_name}: {len(df)} rows")
        except Exception as e:
            print(f"Error loading {path}: {e}")
    return datasets

datasets = load_datasets(file_paths)
```

```
Loaded data_1_1_youth-mortality-rate: 10515 rows
Loaded data_2_2_number-of-infant-deaths-unwpp: 18944 rows
Loaded data_3_3_child-mortality-by-income-level-of-country: 14200 rows
Loaded data_4_4_Distribution of Causes of Death among Children Aged less than 5 years: 146664 rows
Loaded data_5_5_number-of-maternal-deaths-by-region: 7056 rows
Loaded data_6_6_births-attended-by-health-staff-sdgs: 2985 rows
Loaded data_7_7_global-vaccination-coverage: 7897 rows
Loaded data_8_8_health-protection-coverage: 162 rows
```

In [3]:
```python
def clean_dataset(df):
    """
    Comprehensive data cleaning method

    Args:
        df (pd.DataFrame): Input DataFrame

    Returns:
```

```
        pd.DataFrame: Cleaned DataFrame
    """
    # Standardize column names
    df.columns = (df.columns.str.strip()
                  .str.lower()
                  .str.replace(' ', '_')
                  .str.replace('-', '_')
                  .str.replace('[^a-z0-9_]', '', regex=True))

    # Identify numeric columns
    numeric_columns = df.select_dtypes(include=['float64', 'int64']).columns

    # Advanced missing value handling
    for col in numeric_columns:
        # Interpolate for continuous numeric columns
        if df[col].notna().sum() > len(df) * 0.5:  # If more than 50% data is available
            df[col] = df[col].interpolate(method='linear')
        else:
            # Forward fill, then backward fill for sparse data
            df[col] = df[col].ffill().bfill()

    # Remove columns with all missing values
    df = df.dropna(axis=1, how='all')

    return df
```

In [4]:
```
# Clean all datasets
cleaned_datasets = {name: clean_dataset(df) for name, df in datasets.items()}
```

In [5]:
```
# Generate a report
def generate_comprehensive_report(cleaned_datasets):
    """
    Generate a comprehensive report of cleaned datasets

    Args:
        cleaned_datasets (dict): Dictionary of cleaned datasets
    """
    print("\n" + "="*80)
    print("INFANT MORTALITY DATASET - COMPREHENSIVE ANALYSIS REPORT")
    print("="*80 + "\n")

    for name, df in cleaned_datasets.items():
        print(f"DATASET: {name.upper()}")
        print("-"*40)

        # Basic Info
        print(f"Rows: {len(df)}, Columns: {len(df.columns)}")

        # Column Types
        print("\nColumn Types:")
        print(df.dtypes.value_counts())

        # Numeric Column Stats
        numeric_cols = df.select_dtypes(include=['float64', 'int64']).columns
        if len(numeric_cols) > 0:
            print("\nNumeric Column Statistics:")
            print(df[numeric_cols].describe().T)

        # Quick Data Insights
        print("\nQuick Insights:")

        # Year range for datasets with year column
        if 'year' in df.columns:
            print(f"Year Range: {df['year'].min()} - {df['year'].max()}")

        # Unique entities/locations
        if 'entity' in df.columns:
            print(f"Total Unique Locations: {df['entity'].nunique()}")

        print("\n" + "="*40 + "\n")

generate_comprehensive_report(cleaned_datasets)
```

```
================================================================================
INFANT MORTALITY DATASET - COMPREHENSIVE ANALYSIS REPORT
================================================================================

DATASET: DATA_1_1_YOUTH-MORTALITY-RATE
----------------------------------------
Rows: 10515, Columns: 4

Column Types:
object     2
```

```
int64      1
float64    1
Name: count, dtype: int64

Numeric Column Statistics:
                                count         mean        std          min  \
year                            10515.0  1997.837565  16.457225  1950.000000
under_fifteen_mortality_rate    10515.0     6.407323   6.958859     0.190339

                                       25%          50%          75%  \
year                            1988.000000  2000.000000  2011.000000
under_fifteen_mortality_rate       1.670638     3.532842     8.704788

                                       max
year                            2022.000000
under_fifteen_mortality_rate      59.586483

Quick Insights:
Year Range: 1950 - 2022
Total Unique Locations: 232

========================================

DATASET: DATA_2_2_NUMBER-OF-INFANT-DEATHS-UNWPP
----------------------------------------
Rows: 18944, Columns: 4

Column Types:
object     2
int64      1
float64    1
Name: count, dtype: int64

Numeric Column Statistics:
                                                  count           mean  \
year                                            18944.0    1986.500000
deaths___sex_all___age_0___variant_estimates    18944.0  246156.674416

                                                        std     min     25%  \
year                                           2.136057e+01  1950.0  1968.0
deaths___sex_all___age_0___variant_estimates   1.198653e+06     0.0   184.0

                                                  50%       75%         max
year                                           1986.5   2005.00      2023.0
deaths___sex_all___age_0___variant_estimates   3023.0  28610.25  13514505.0

Quick Insights:
Year Range: 1950 - 2023
Total Unique Locations: 256

========================================

DATASET: DATA_3_3_CHILD-MORTALITY-BY-INCOME-LEVEL-OF-COUNTRY
----------------------------------------
Rows: 14200, Columns: 4

Column Types:
object     2
int64      1
float64    1
Name: count, dtype: int64

Numeric Column Statistics:
                                                    count         mean  \
year                                              14200.0  1989.947042
observation_value___indicator_under_five_mortal...  14200.0     8.009897

                                                         std          min  \
year                                               20.417969  1932.000000
observation_value___indicator_under_five_mortal...   8.409812     0.146058

                                                           25%         50%  \
year                                               1974.000000  1992.00000
observation_value___indicator_under_five_mortal...     1.811918     4.70367

                                                           75%         max
year                                               2007.000000  2022.00000
observation_value___indicator_under_five_mortal...    11.561593    57.14969

Quick Insights:
Year Range: 1932 - 2022
Total Unique Locations: 232
```

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

DATASET: DATA_4_4_DISTRIBUTION OF CAUSES OF DEATH AMONG CHILDREN AGED LESS THAN 5 YEARS
----------------------------------------
Rows: 146664, Columns: 21

Column Types:
object     17
float64     2
int64       1
bool        1
Name: count, dtype: int64

Numeric Column Statistics:
                    count         mean       std     min        25%      50%  \
period            146664.0  2008.500000  5.188145  2000.0  2004.0000  2008.50
factvaluenumeric  146664.0     0.071072  0.097356     0.0     0.0001     0.03
value             146664.0     0.068298  0.102069     0.0     0.0000     0.00

                     75%     max
period            2013.0  2017.0
factvaluenumeric     0.1     1.0
value                0.1     1.0

Quick Insights:

========================================

DATASET: DATA_5_5_NUMBER-OF-MATERNAL-DEATHS-BY-REGION
----------------------------------------
Rows: 7056, Columns: 5

Column Types:
object     3
int64      1
float64    1
Name: count, dtype: int64

Numeric Column Statistics:
                            count         mean           std          min  \
year                       7056.0  2002.500000     10.389031  1985.000000
estimated_maternal_deaths  7056.0  8661.259033  43820.662340     0.061462

                                  25%          50%        75%       max
year                       1993.750000  2002.500000  2011.2500    2020.0
estimated_maternal_deaths     9.854105   127.637903  1664.7789  578245.9

Quick Insights:
Year Range: 1985 - 2020
Total Unique Locations: 196

========================================

DATASET: DATA_6_6_BIRTHS-ATTENDED-BY-HEALTH-STAFF-SDGS
----------------------------------------
Rows: 2985, Columns: 4

Column Types:
object     2
int64      1
float64    1
Name: count, dtype: int64

Numeric Column Statistics:
                                                    count         mean  \
year                                               2985.0  2005.550419
births_attended_by_skilled_health_staff__of_total  2985.0    88.926417

                                                         std     min     25%  \
year                                                8.543703  1980.0  1999.0
births_attended_by_skilled_health_staff__of_total  20.156612     5.0    90.0

                                                      50%      75%     max
year                                               2006.0  2013.00  2021.0
births_attended_by_skilled_health_staff__of_total    98.7    99.78   100.0

Quick Insights:
Year Range: 1980 - 2021
Total Unique Locations: 219

========================================

DATASET: DATA_7_7_GLOBAL-VACCINATION-COVERAGE
```

```
----------------------------------------
Rows: 7897, Columns: 14

Column Types:
float64    11
object      2
int64       1
Name: count, dtype: int64

Numeric Column Statistics:
                                    count         mean        std     min  \
year                                7897.0  2001.586299  11.769320  1980.0
bcg__of_one_year_olds_immunized     7897.0    81.314170  21.512878     0.0
hepb3__of_one_year_olds_immunized   7872.0    75.828887  23.074306     0.0
hib3__of_one_year_olds_immunized    7897.0    82.793846  19.403296     0.0
ipv1__of_one_year_olds_immunized    7897.0    85.691402  16.897581     0.0
mcv1__of_one_year_olds_immunized    7897.0    77.999683  22.099107     0.0
pcv3__of_one_year_olds_immunized    7897.0    77.940864  22.681078     0.0
pol3__of_one_year_olds_immunized    7897.0    80.016019  22.149832     0.0
rcv1__of_one_year_olds_immunized    7855.0    83.846913  18.992296     0.0
rotac__of_one_year_olds_immunized   7897.0    72.222236  22.698781     0.0
yfv__of_one_year_olds_immunized     7897.0    61.097759  26.122419     0.0
dtp3__of_one_year_olds_immunized    7897.0    79.462897  22.248666     1.0

                                           25%          50%          75%  \
year                                1992.00000  2002.000000  2012.000000
bcg__of_one_year_olds_immunized       73.00000    90.000000    97.000000
hepb3__of_one_year_olds_immunized     65.00000    83.410526    94.000000
hib3__of_one_year_olds_immunized      77.00000    90.000000    96.000000
ipv1__of_one_year_olds_immunized      79.00000    93.000000    98.000000
mcv1__of_one_year_olds_immunized      67.00000    86.000000    95.000000
pcv3__of_one_year_olds_immunized      70.00000    85.000000    94.000000
pol3__of_one_year_olds_immunized      72.00000    89.000000    96.000000
rcv1__of_one_year_olds_immunized      80.20943    90.000000    95.968085
rotac__of_one_year_olds_immunized     61.00000    80.000000    87.000000
yfv__of_one_year_olds_immunized       53.00000    64.000000    85.000000
dtp3__of_one_year_olds_immunized      71.00000    88.000000    96.000000

                                      max
year                                2021.0
bcg__of_one_year_olds_immunized       99.0
hepb3__of_one_year_olds_immunized     99.0
hib3__of_one_year_olds_immunized      99.0
ipv1__of_one_year_olds_immunized      99.0
mcv1__of_one_year_olds_immunized      99.0
pcv3__of_one_year_olds_immunized      99.0
pol3__of_one_year_olds_immunized      99.0
rcv1__of_one_year_olds_immunized      99.0
rotac__of_one_year_olds_immunized     99.0
yfv__of_one_year_olds_immunized       99.0
dtp3__of_one_year_olds_immunized      99.0

Quick Insights:
Year Range: 1980 - 2021
Total Unique Locations: 202

========================================

DATASET: DATA_8_8_HEALTH-PROTECTION-COVERAGE
----------------------------------------
Rows: 162, Columns: 4

Column Types:
object     2
int64      1
float64    1
Name: count, dtype: int64

Numeric Column Statistics:
                                                    count         mean  \
year                                                162.0  2008.376543
share_of_population_covered_by_health_insurance...  162.0    62.507407

                                                          std     min     25%  \
year                                                 2.702691  1995.0  2008.0
share_of_population_covered_by_health_insurance...  39.557004     0.0    23.0

                                                       50%     75%     max
year                                                2009.0  2010.0  2011.0
share_of_population_covered_by_health_insurance...    83.5   100.0   100.0

Quick Insights:
Year Range: 1995 - 2011
```

```
Total Unique Locations: 162

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

Overview of Datasets

The analysis covers 8 different datasets focusing on various aspects of child and maternal health:

1. Youth Mortality Rate
2. Number of Infant Deaths
3. Child Mortality by Income Level
4. Distribution of Causes of Death (Children Under 5)
5. Maternal Deaths by Region
6. Births Attended by Health Staff
7. Global Vaccination Coverage
8. Health Protection Coverage

Key Observations

1. Mortality Rates

- **Under-15 Mortality Rate** (1950-2022):
  - Mean: 6.41 deaths per 1,000
  - Wide variation: Ranges from 0.19 to 59.59
  - Median: 3.53 deaths per 1,000

2. Infant and Child Deaths

- **Number of Infant Deaths** (1950-2023):
  - Mean: 246,157 deaths per year
  - Significant variation (std dev: 1,198,653)
  - Maximum recorded: 13,514,505 deaths in a single year/location

3. Under-5 Mortality

- **Global Under-5 Mortality Rate** (1932-2022):
  - Mean: 8.01 deaths per 1,000
  - Range: 0.15 to 57.15
  - Median: 4.70 deaths per 1,000

4. Vaccination Coverage

Average vaccination rates for one-year-olds (1980-2021):

- BCG: 81.3%
- Hepatitis B3: 75.8%
- Hib3: 82.8%
- Polio3: 80.0%
- Measles: 78.0%
- DTP3: 79.5%

5. Health Coverage

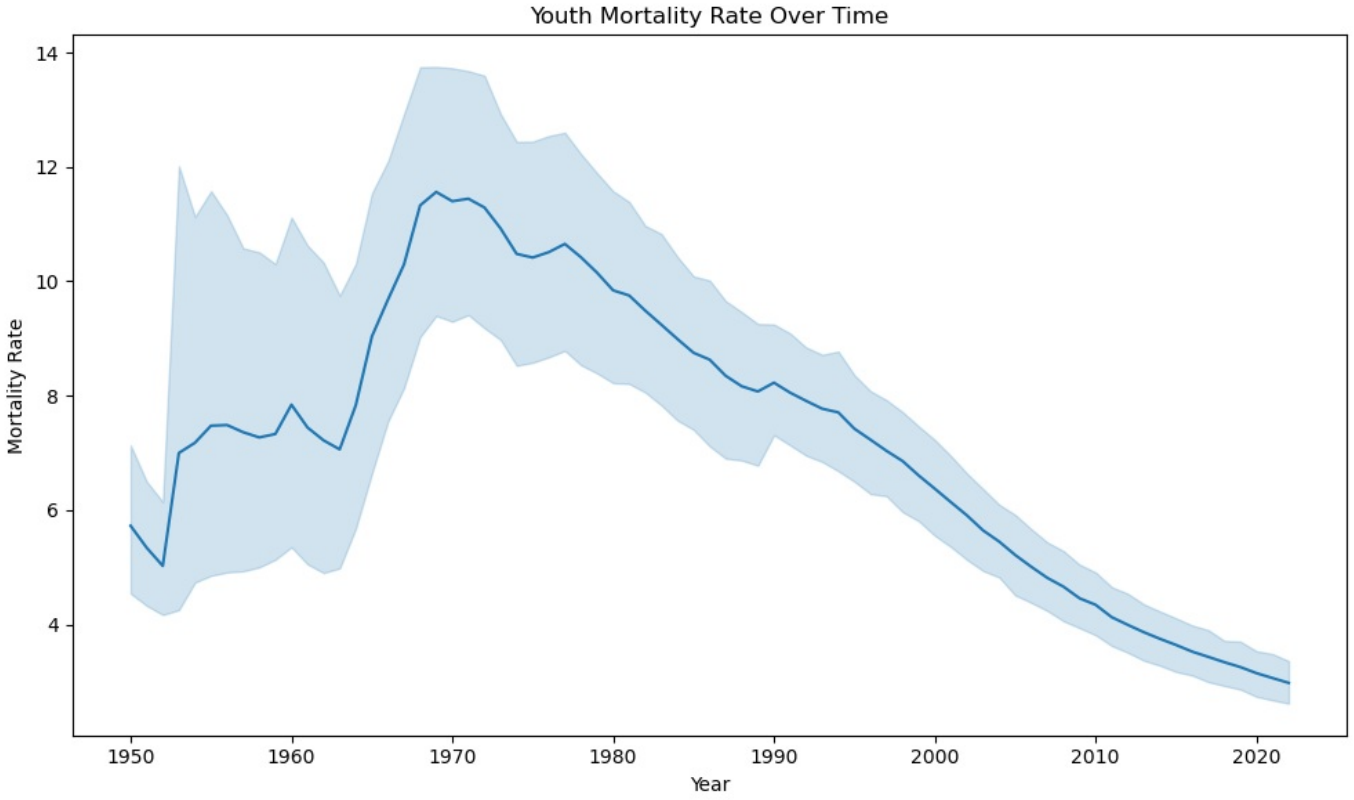
- **Births Attended by Skilled Health Staff** (1980-2021):
  - Mean: 88.9%
  - Median: 98.7%
  - Suggesting significant improvement in medical care

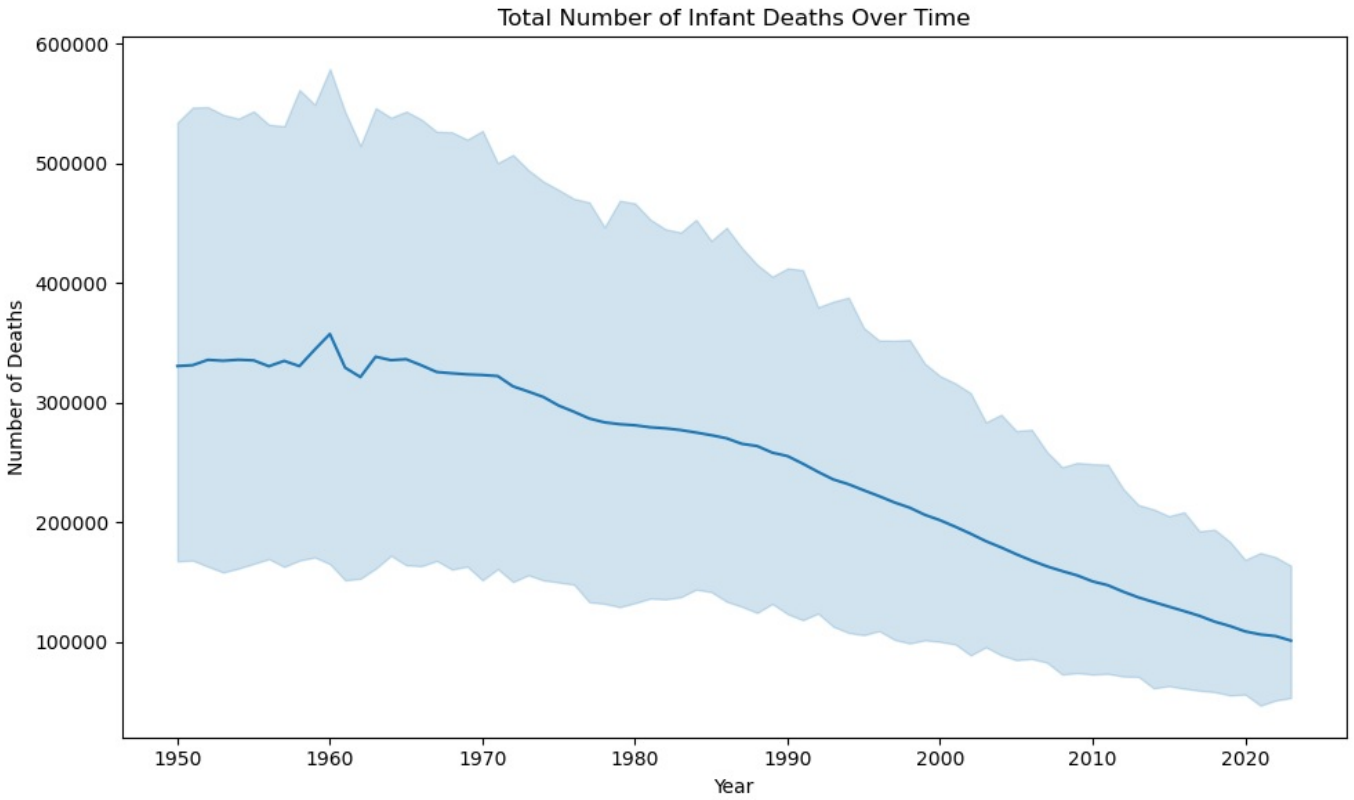
Data Visualization

In [6]:
```
# 1. Youth Mortality Rate Trend
plt.figure(figsize=(10, 6))
mortality_data = cleaned_datasets['data_1_1_youth-mortality-rate']
try:
    sns.lineplot(data=mortality_data, x='year', y='under_fifteen_mortality_rate')
    plt.title('Youth Mortality Rate Over Time')
    plt.xlabel('Year')
```

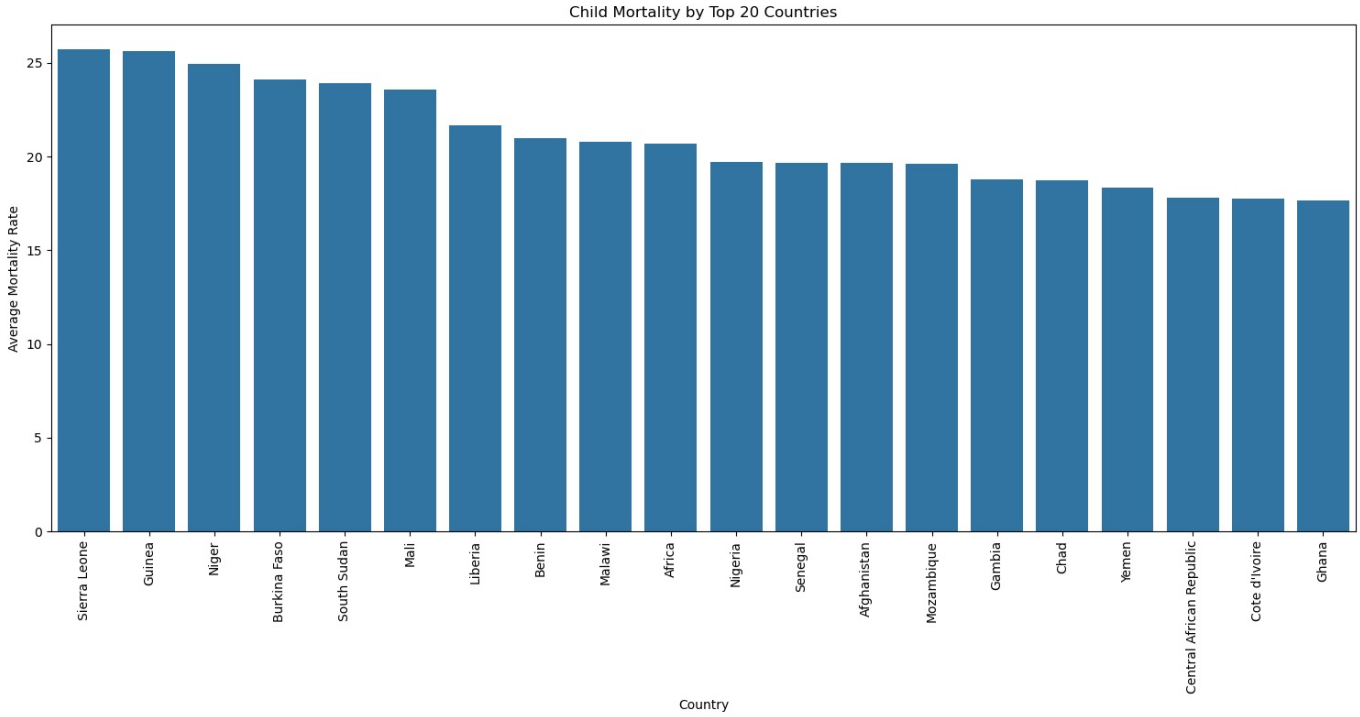
```
    plt.ylabel('Mortality Rate')
    plt.tight_layout()
    plt.show()
except Exception as e:
    print(f"Error in Mortality Rate Plot: {str(e)}")

# 2. Total Infant Deaths Trend
plt.figure(figsize=(10, 6))
infant_deaths_data = cleaned_datasets['data_2_2_number-of-infant-deaths-unwpp']
try:
    sns.lineplot(data=infant_deaths_data, x='year',
                 y='deaths___sex_all___age_0___variant_estimates')
    plt.title('Total Number of Infant Deaths Over Time')
    plt.xlabel('Year')
    plt.ylabel('Number of Deaths')
    plt.tight_layout()
    plt.show()
except Exception as e:
    print(f"Error in Infant Deaths Plot: {str(e)}")

# 3. Child Mortality by Country
plt.figure(figsize=(15, 8))
child_mortality_data = cleaned_datasets['data_3_3_child-mortality-by-income-level-of-country']
try:
    # Find appropriate mortality column dynamically
    mortality_columns = [col for col in child_mortality_data.columns
                         if 'mortality' in col.lower()]

    if mortality_columns:
        mortality_column = mortality_columns[0]

        # Aggregate data for boxplot
        plot_data = child_mortality_data.groupby('entity')[mortality_column].mean().reset_index()

        # Sort and select top 20 countries for readability
        plot_data = plot_data.nlargest(20, mortality_column)

        plt.figure(figsize=(15, 8))
        sns.barplot(x='entity', y=mortality_column, data=plot_data)
        plt.title('Child Mortality by Top 20 Countries')
        plt.xlabel('Country')
        plt.ylabel('Average Mortality Rate')
        plt.xticks(rotation=90)
        plt.tight_layout()
        plt.show()
    else:
        print("No mortality column found")
except Exception as e:
    print(f"Error in Child Mortality Plot: {str(e)}")

# 4. Vaccination Coverage
plt.figure(figsize=(12, 6))
vaccination_data = cleaned_datasets['data_7_7_global-vaccination-coverage']
try:
    # Find vaccination columns
    vacc_columns = [col for col in vaccination_data.columns
                    if 'immunized' in col.lower()]

    plt.title('Vaccination Coverage Over Time')
    for column in vacc_columns[:3]:  # Limit to first 3 for clarity
        plt.plot(vaccination_data['year'], vaccination_data[column], label=column)

    plt.xlabel('Year')
    plt.ylabel('Immunization Coverage (%)')
    plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
    plt.tight_layout()
    plt.show()
except Exception as e:
    print(f"Error in Vaccination Coverage Plot: {str(e)}")

# 5. Health Coverage
plt.figure(figsize=(10, 6))
health_coverage_data = cleaned_datasets['data_8_8_health-protection-coverage']
try:
    # Identify numeric columns excluding 'year'
    numeric_columns = health_coverage_data.select_dtypes(include=[np.number]).columns
    coverage_column = [col for col in numeric_columns if col != 'year'][0]

    plt.plot(health_coverage_data['year'], health_coverage_data[coverage_column])
    plt.title('Health Insurance Coverage Over Time')
    plt.xlabel('Year')
    plt.ylabel('Coverage Percentage')
    plt.tight_layout()
```

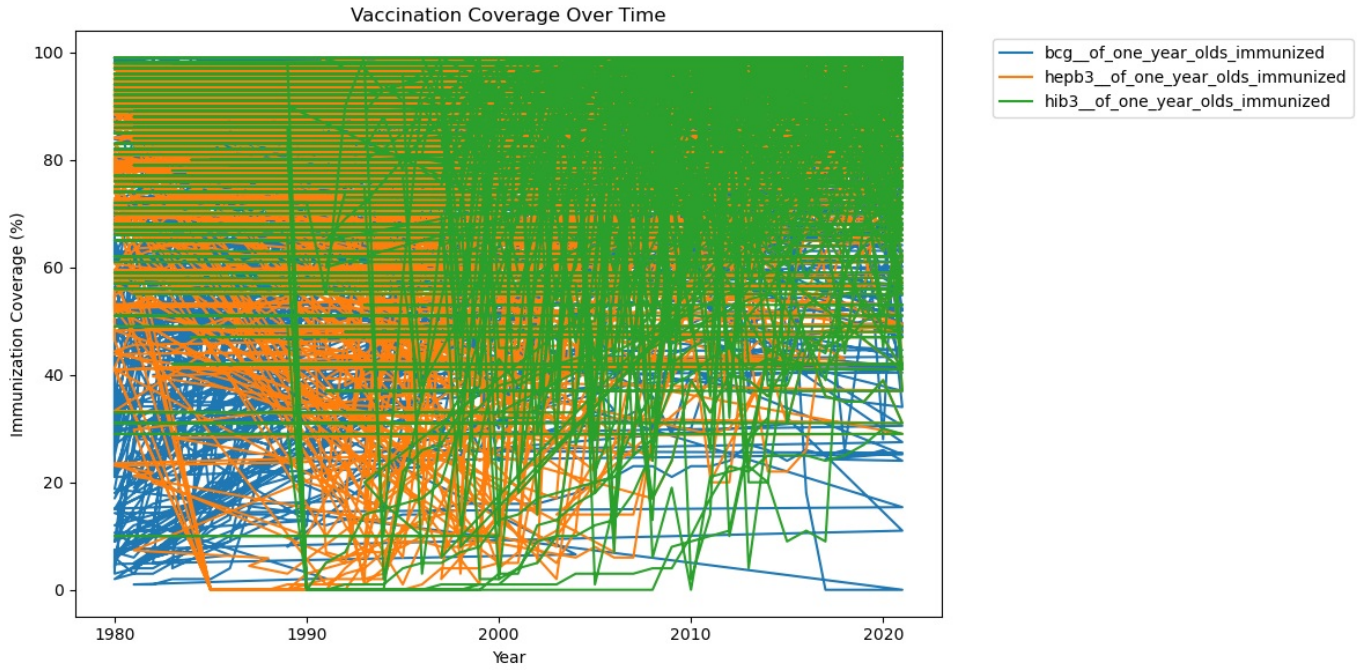
```
    plt.show()
except Exception as e:
    print(f"Error in Health Coverage Plot: {str(e)}")
```

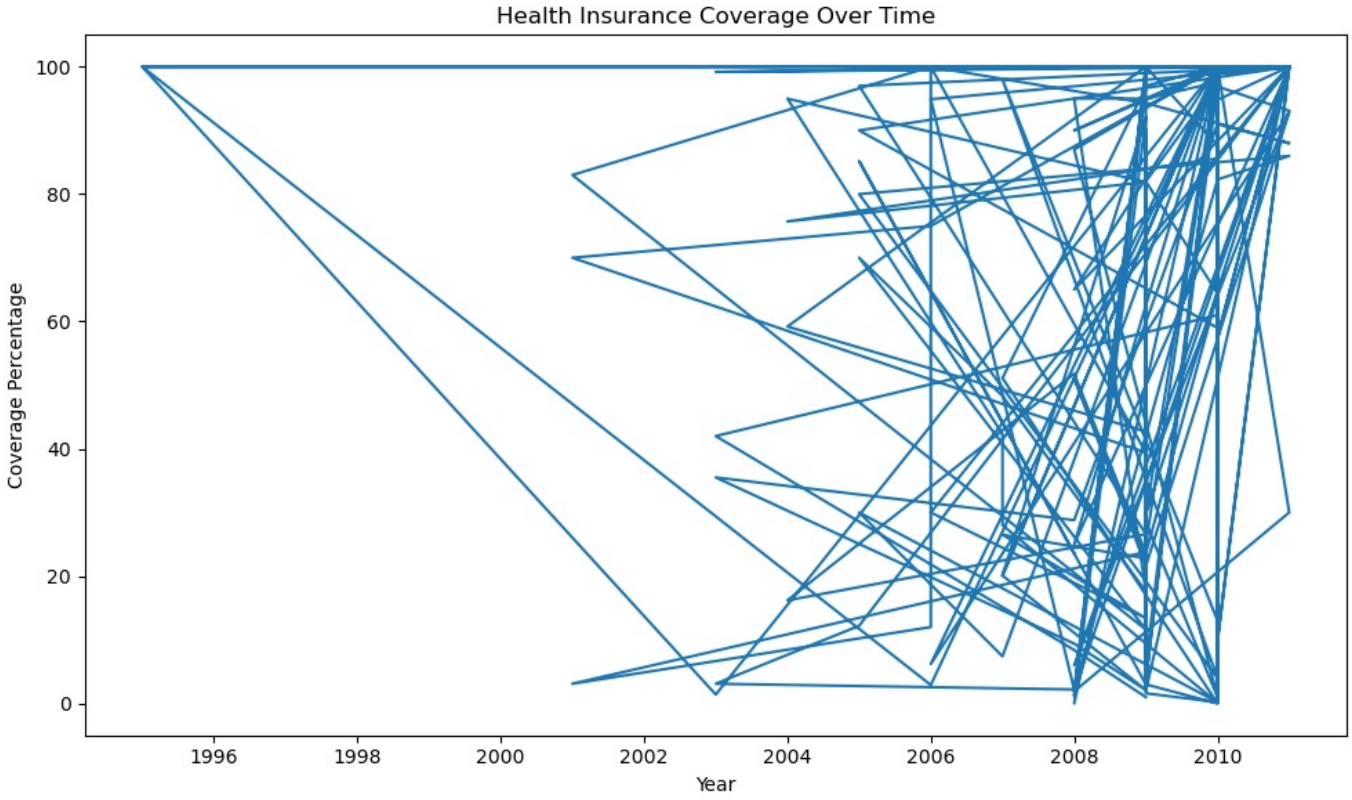





```
<Figure size 1500x800 with 0 Axes>
```


Child Mortality by Top 20 Countries
Average Mortality Rate
0
5
10
15
20
25
Sierra Leone
Guinea
Niger
Burkina Faso
South Sudan
Mali
Liberia
Benin
Malawi
Africa
Nigeria
Senegal
Afghanistan
Mozambique
Gambia
Chad
Yemen
Central African Republic
Cote d'Ivoire
Ghana
Country


Vaccination Coverage Over Time
Immunization Coverage (%)
0
20
40
60
80
100
1980
1990
2000
2010
2020
Year
bcg__of_one_year_olds_immunized
hepb3__of_one_year_olds_immunized
hib3__of_one_year_olds_immunized



In [ ]: