Health Investments and Outcomes: Exploring Global Patterns in Expenditure, Sanitation, and Life Expectancy

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
from datetime import datetime
```

Loading and Accessing Datasets

In [37]:
```
sani = pd.read_excel("C:\\Users\\navya\\Downloads\\Python_Project\\sanitation data.
ghed = pd.read_excel("C:\\Users\\navya\\Downloads\\Python_Project\\GHED all data (A
expec = pd.read_excel("C:\\Users\\navya\\Downloads\\Python_Project\\life expectancy
```

Cleaning Columns and rows

In [38]:
```
def clean_columns(df):
    df.columns = df.columns.str.strip().str.lower()
    return df

ghed = clean_columns(ghed)
expec = clean_columns(expec)
sani  = clean_columns(sani)
```

In [39]:
```
ghed
```

Out[39]:

	location	code	region	income	year	pop_size	che_gdp	che_pc_usd	che
0	Algeria	DZA	AFR	Lower-middle	2000	1	3.214854	61.857853	1.438703e+05
1	Algeria	DZA	AFR	Lower-middle	2001	1	3.536286	67.058594	1.622309e+05
2	Algeria	DZA	AFR	Lower-middle	2002	1	3.441696	66.681633	1.687023e+05
3	Algeria	DZA	AFR	Lower-middle	2003	1	3.325694	75.951309	1.891375e+05
4	Algeria	DZA	AFR	Lower-middle	2004	1	3.290305	92.687630	2.179286e+05
...	...	...	...	...	...	...	...	...	...
4401	Viet Nam	VNM	WPR	Lower-middle	2018	1	5.026788	161.978714	3.523297e+08
4402	Viet Nam	VNM	WPR	Lower-middle	2019	1	4.974075	171.152969	3.833619e+08
4403	Viet Nam	VNM	WPR	Lower-middle	2020	1	4.332198	153.101608	3.484988e+08
4404	Viet Nam	VNM	WPR	Lower-middle	2021	1	4.537569	168.080338	3.851251e+08
4405	Viet Nam	VNM	WPR	Lower-middle	2022	1	4.588916	188.897491	4.381836e+08

4406 rows × 4121 columns

In [40]:
```
expec
```

Out[40]:

	ind_id	ind_code	ind_uuid	ind_per_code	dim_time	dim_t
0	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2000	
1	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2000	
2	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2000	
3	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2000	
4	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2000	
...	...	...	...	...	...	
12931	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2021	
12932	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2021	
12933	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2021	
12934	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2021	
12935	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2021	

12936 rows × 15 columns

In [41]:
```
sani
```

Out[41]:

	data source	world development indicators	unnamed: 2	unnamed: 3	unnamed: 4	unnamed: 5	unnamed: 6
0	Last Updated Date	2025-10-07 00:00:00	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Country Name	Country Code	Indicator Name	Indicator Code	1960.0	1961.0	1962.0
3	Aruba	ABW	People using at least basic sanitation service...	SH.STA.BASS.ZS	NaN	NaN	NaN
4	Africa Eastern and Southern	AFE	People using at least basic sanitation service...	SH.STA.BASS.ZS	NaN	NaN	NaN
...	...	...	...	...	...	...	...
264	Kosovo	XKX	People using at least basic sanitation service...	SH.STA.BASS.ZS	NaN	NaN	NaN
265	Yemen, Rep.	YEM	People using at least basic sanitation service...	SH.STA.BASS.ZS	NaN	NaN	NaN
266	South Africa	ZAF	People using at least basic sanitation service...	SH.STA.BASS.ZS	NaN	NaN	NaN
267	Zambia	ZMB	People using at least basic sanitation service...	SH.STA.BASS.ZS	NaN	NaN	NaN
268	Zimbabwe	ZWE	People using at least basic sanitation service...	SH.STA.BASS.ZS	NaN	NaN	NaN

269 rows × 69 columns

In [42]:
```
print("GHED columns:", ghed.columns.tolist()[:10])
print("EXPEC columns:", expec.columns.tolist()[:10])
print("SANITATION columns:", sani.columns.tolist()[:10])
```

```
GHED columns: ['location', 'code', 'region', 'income', 'year', 'pop_size', 'che_gd
p', 'che_pc_usd', 'che', 'gghed']
EXPEC columns: ['ind_id', 'ind_code', 'ind_uuid', 'ind_per_code', 'dim_time', 'dim_t
ime_type', 'dim_geo_code_m49', 'dim_geo_code_type', 'dim_publish_state_code', 'ind_n
ame']
SANITATION columns: ['data source', 'world development indicators', 'unnamed: 2', 'u
nnamed: 3', 'unnamed: 4', 'unnamed: 5', 'unnamed: 6', 'unnamed: 7', 'unnamed: 8', 'u
nnamed: 9']
```

Preprocessing with ghed

In [43]:
```
ghed.isnull()
```

Out[43]:

	location	code	region	income	year	pop_size	che_gdp	che_pc_usd	che	gghed
0	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...
4401	False	False	False	False	False	False	False	False	False	False
4402	False	False	False	False	False	False	False	False	False	False
4403	False	False	False	False	False	False	False	False	False	False
4404	False	False	False	False	False	False	False	False	False	False
4405	False	False	False	False	False	False	False	False	False	False

4406 rows × 4121 columns

In [44]:
```
ghed.isnull().sum()
```

Out[44]:
```
location                0
code                    0
region                  0
income                  0
year                    0
                      ...
hk_ext_usd2022_pc    3647
hk_pvtd_usd2022_pc   4029
gdp_usd2022_pc         23
pfc_usd2022_pc         28
gge_usd2022_pc         23
Length: 4121, dtype: int64
```

In [45]:
```
ghed.duplicated().sum()
```

Out[45]:
```
0
```

with sanitation data

In [46]:
```
sani.isna()
```

Out[46]:

	data source	world development indicators	unnamed: 2	unnamed: 3	unnamed: 4	unnamed: 5	unnamed: 6	unnam
0	False	False	True	True	True	True	True	
1	True	True	True	True	True	True	True	
2	False	False	False	False	False	False	False	F
3	False	False	False	False	True	True	True	
4	False	False	False	False	True	True	True	
...	...	...	...	...	...	...	...	
264	False	False	False	False	True	True	True	
265	False	False	False	False	True	True	True	
266	False	False	False	False	True	True	True	
267	False	False	False	False	True	True	True	
268	False	False	False	False	True	True	True	

269 rows × 69 columns

In [47]:
```
sani.isnull()
```

Out[47]:

	data source	world development indicators	unnamed: 2	unnamed: 3	unnamed: 4	unnamed: 5	unnamed: 6	unnam
0	False	False	True	True	True	True	True	
1	True	True	True	True	True	True	True	
2	False	False	False	False	False	False	False	F
3	False	False	False	False	True	True	True	
4	False	False	False	False	True	True	True	
...	...	...	...	...	...	...	...	
264	False	False	False	False	True	True	True	
265	False	False	False	False	True	True	True	
266	False	False	False	False	True	True	True	
267	False	False	False	False	True	True	True	
268	False	False	False	False	True	True	True	

269 rows × 69 columns

In [48]:
```
sani.isnull().sum()
```

Out[48]:
```
data source                     1
world development indicators    1
unnamed: 2                      2
unnamed: 3                      2
unnamed: 4                    268
                             ...
unnamed: 64                    21
unnamed: 65                    24
unnamed: 66                    30
unnamed: 67                   268
unnamed: 68                   268
Length: 69, dtype: int64
```

In [49]:
```
sani.duplicated().sum()
```

Out[49]:
```
0
```

Merging

In [56]:
```
expec = expec.copy()
ghed  = ghed.copy()
```

In [57]:
```
expec['country'] = expec['country'].astype(str).str.strip()
ghed['country']  = ghed['location'].astype(str).str.strip()
```

In [58]:
```
expec['year'] = pd.to_numeric(expec['year'], errors='coerce').astype('Int64')
ghed['year']  = pd.to_numeric(ghed['year'], errors='coerce').astype('Int64')
```

In [59]:
```
print("life: unique countries", expec['country'].nunique(), "years", expec['year'].
print("ghed: unique countries", ghed['country'].nunique(), "years", ghed['year'].nu
```

```
life: unique countries 196 years 22
ghed: unique countries 194 years 24
```

In [64]:
```
merged_tmp = pd.merge(
    expec,
    ghed.drop(columns=['location'], errors='ignore'),
    on=['country','year'],
    how='outer',
    indicator=True
)

print("\nMerge indicator counts:\n", merged_tmp['_merge'].value_counts())
```

```
Merge indicator counts:
 _merge
both          11715
left_only      1221
right_only      501
Name: count, dtype: int64
```

In [65]:
```
merged_both = merged_tmp[merged_tmp['_merge']=='both'].drop(columns=['_merge']).cop
print("\nMerged (both) shape:", merged_both.shape)
display(merged_both.head())
```

```
Merged (both) shape: (11715, 4134)
```

	ind_id	ind_code	ind_uuid	ind_per_code	year	dim_time_type
6	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2002	YEAR
7	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2002	YEAR
8	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2002	YEAR
9	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2003	YEAR
10	90E2E48WHOSIS_000001	WHOSIS_000001	90E2E48	WHOSIS_000001	2003	YEAR

5 rows × 4134 columns

In [94]:
```
clean_df = final_df[keep_cols].dropna()
print("Clean dataset shape:", clean_df.shape)
clean_df.head()
```

```
Clean dataset shape: (9834, 8)
```

Out[94]:

	country	year	life_expectancy	sanitation	che_gdp	che_pc_usd	gghed_gdp	oop_pc
0	Afghanistan	2002	55.401586	22.541339	9.443391	16.706974	0.084181	14.26
1	Afghanistan	2002	54.921993	22.541339	9.443391	16.706974	0.084181	14.26
2	Afghanistan	2002	55.154478	22.541339	9.443391	16.706974	0.084181	14.26
3	Afghanistan	2003	56.242837	24.100333	8.941258	17.746025	0.650963	15.27
4	Afghanistan	2003	55.949458	24.100333	8.941258	17.746025	0.650963	15.27

In [96]:
```
clean_df
```

Out[96]:

	country	year	life_expectancy	sanitation	che_gdp	che_pc_usd	gghed_gdp	oo
0	Afghanistan	2002	55.401586	22.541339	9.443391	16.706974	0.084181	1
1	Afghanistan	2002	54.921993	22.541339	9.443391	16.706974	0.084181	1
2	Afghanistan	2002	55.154478	22.541339	9.443391	16.706974	0.084181	1
3	Afghanistan	2003	56.242837	24.100333	8.941258	17.746025	0.650963	1
4	Afghanistan	2003	55.949458	24.100333	8.941258	17.746025	0.650963	1
...	...	...	...	...	...	...	...	
9832	Zimbabwe	2020	57.112639	35.192362	2.954401	51.142506	0.652689	
9833	Zimbabwe	2020	59.404763	35.192362	2.954401	51.142506	0.652689	
9834	Zimbabwe	2021	60.533378	34.609950	2.785717	63.511448	0.912499	
9835	Zimbabwe	2021	56.194609	34.609950	2.785717	63.511448	0.912499	
9836	Zimbabwe	2021	58.481022	34.609950	2.785717	63.511448	0.912499	

9834 rows × 8 columns

Data Exploration and Visulaisation insights

In [97]:

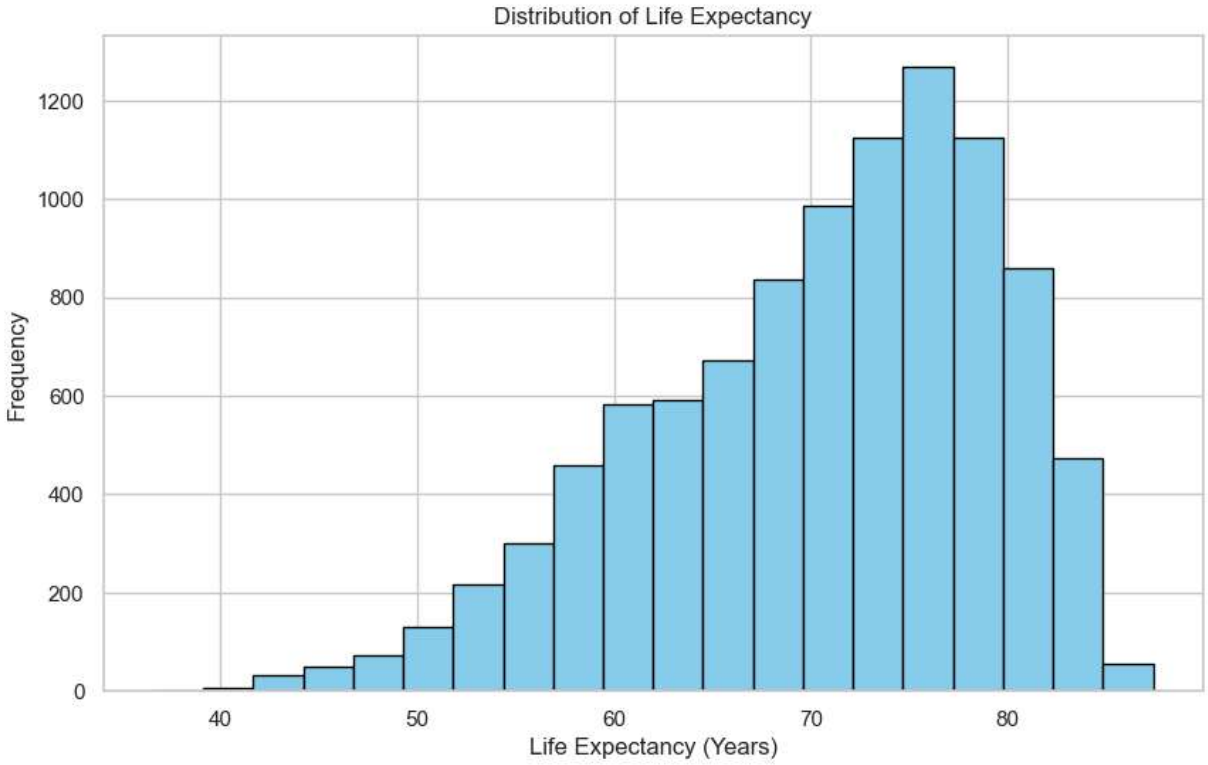
```
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns

X = clean_df[['sanitation', 'che_gdp']]
y = clean_df['life_expectancy']

X = sm.add_constant(X)
model = sm.OLS(y, X).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        life_expectancy   R-squared:                       0.652
Model:                            OLS   Adj. R-squared:                  0.652
Method:                 Least Squares   F-statistic:                     9209.
Date:                Wed, 12 Nov 2025   Prob (F-statistic):               0.00
Time:                        01:21:01   Log-Likelihood:                -30124.
No. Observations:                9834   AIC:                         6.025e+04
Df Residuals:                    9831   BIC:                         6.028e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         52.4146      0.163    320.940      0.000      52.094      52.735
sanitation     0.2245      0.002    123.983      0.000       0.221       0.228
che_gdp        0.3389      0.022     15.554      0.000       0.296       0.382
==============================================================================
Omnibus:                     1390.967   Durbin-Watson:                   0.582
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2655.753
Skew:                          -0.894   Prob(JB):                         0.00
Kurtosis:                       4.812   Cond. No.                         244.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
```

In [99]:
```
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style="whitegrid")
```

Histogram of life expectancy

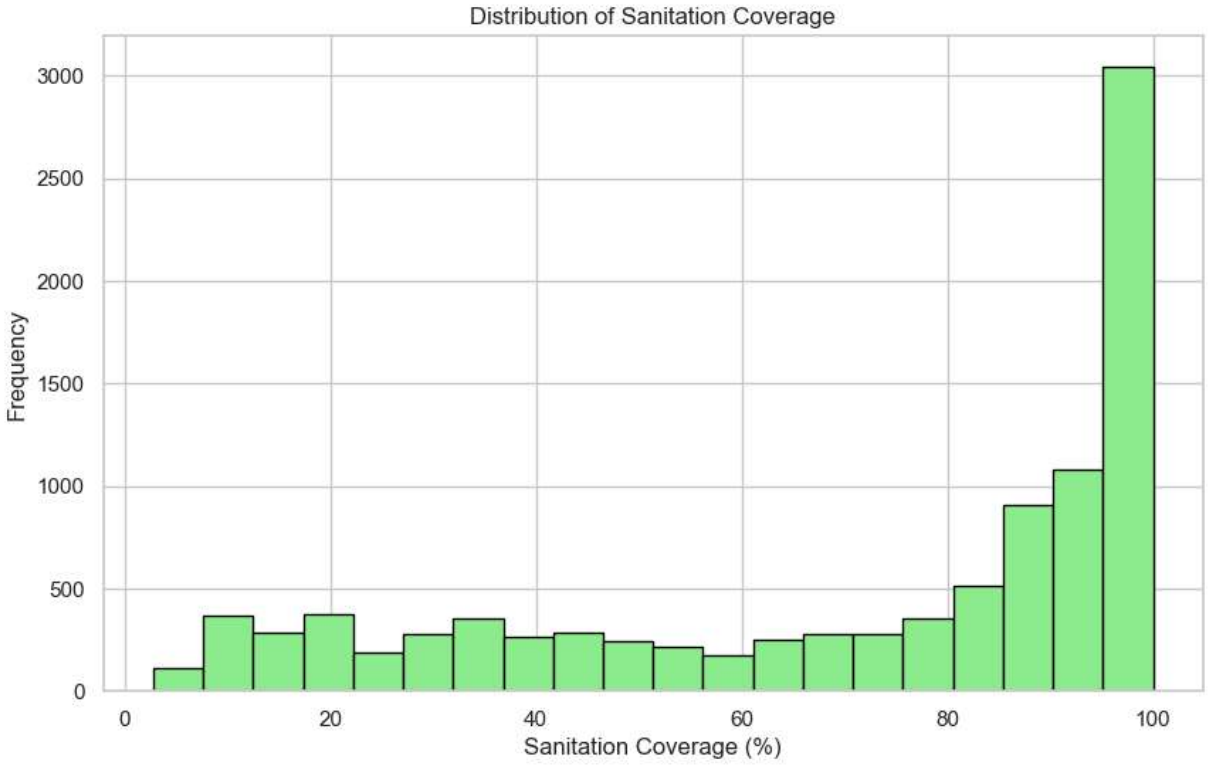
In [100…
```
plt.figure(figsize=(10,6))
plt.hist(clean_df['life_expectancy'], bins=20, color='skyblue', edgecolor='black')
plt.title('Distribution of Life Expectancy')
plt.xlabel('Life Expectancy (Years)')
plt.ylabel('Frequency')
plt.show()
```



Histogram of sanitation data

In [101…

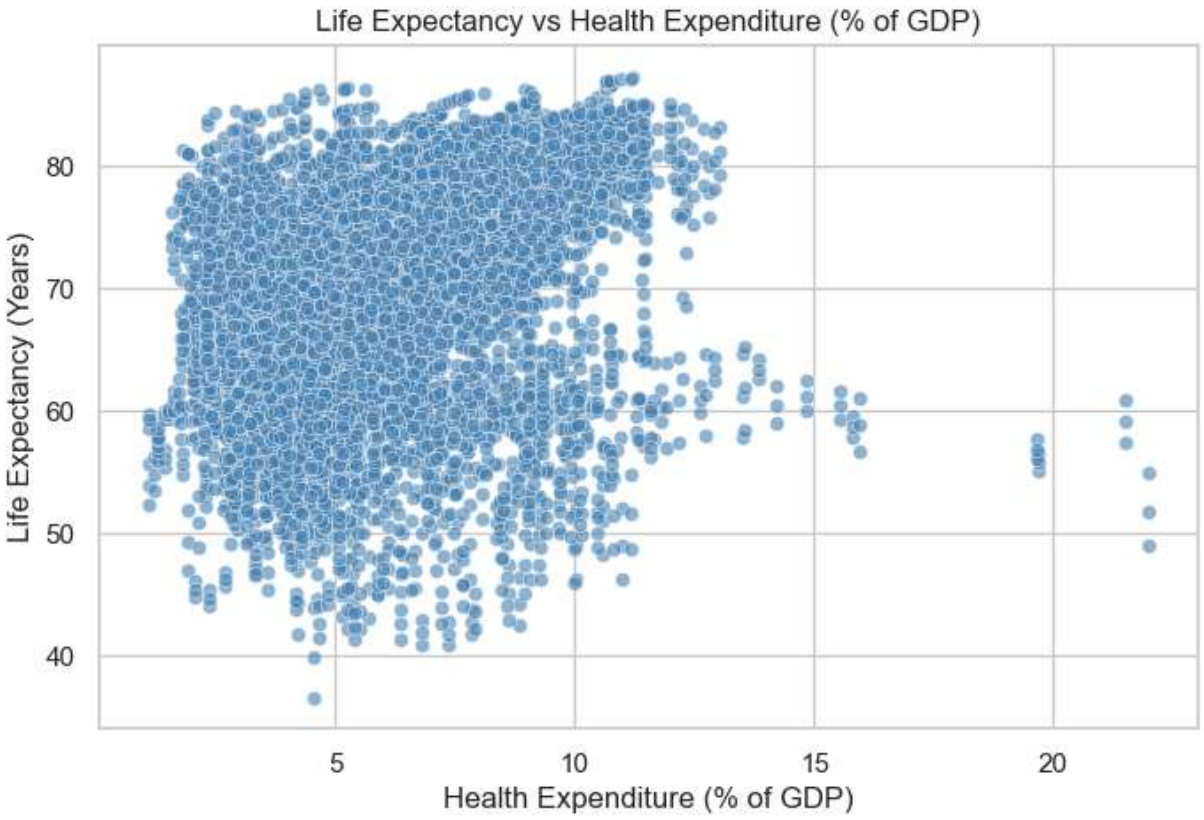
```
plt.figure(figsize=(10,6))
plt.hist(clean_df['sanitation'], bins=20, color='lightgreen', edgecolor='black')
plt.title('Distribution of Sanitation Coverage')
plt.xlabel('Sanitation Coverage (%)')
plt.ylabel('Frequency')
plt.show()
```



Scatterplot life expectancy vs health experience

In [102…

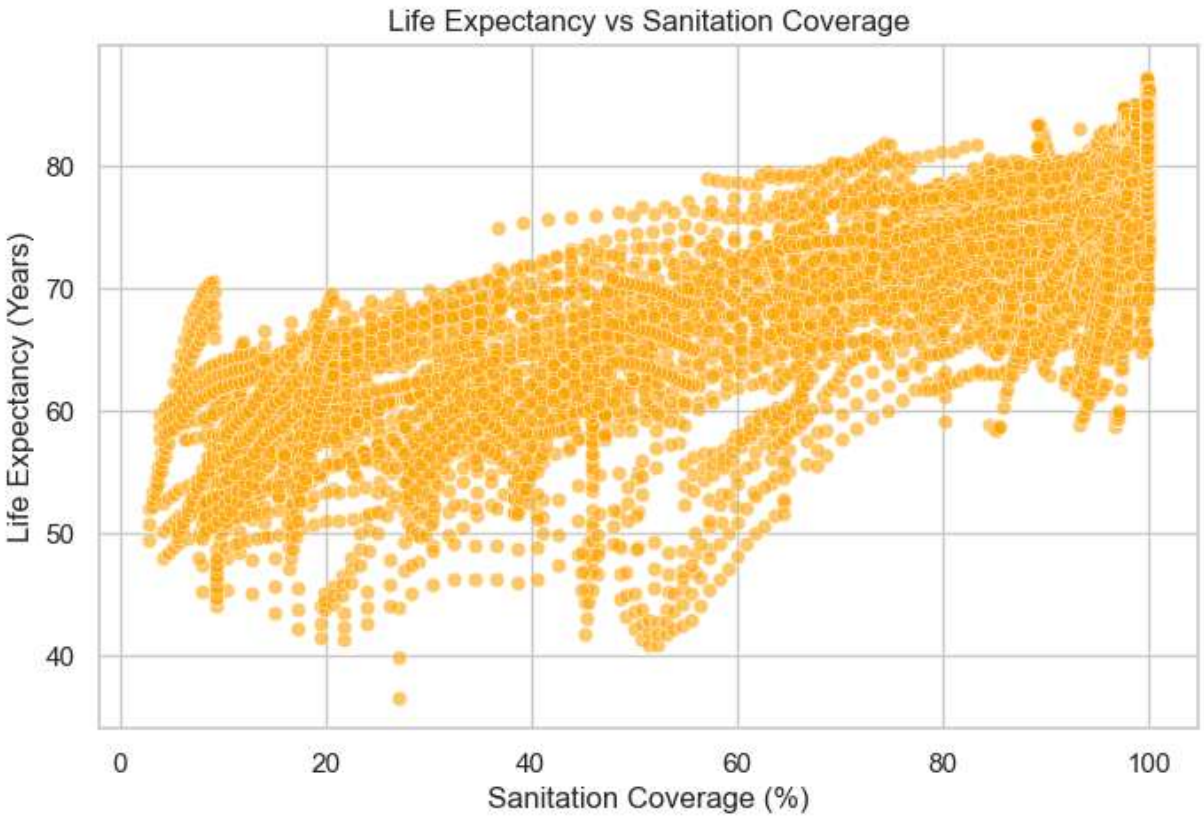
```
plt.figure(figsize=(8,5))
sns.scatterplot(data=clean_df, x='che_gdp', y='life_expectancy', alpha=0.6, color='
plt.title('Life Expectancy vs Health Expenditure (% of GDP)')
plt.xlabel('Health Expenditure (% of GDP)')
plt.ylabel('Life Expectancy (Years)')
plt.show()
```



Life expentancy vs sanitation data

In [103…

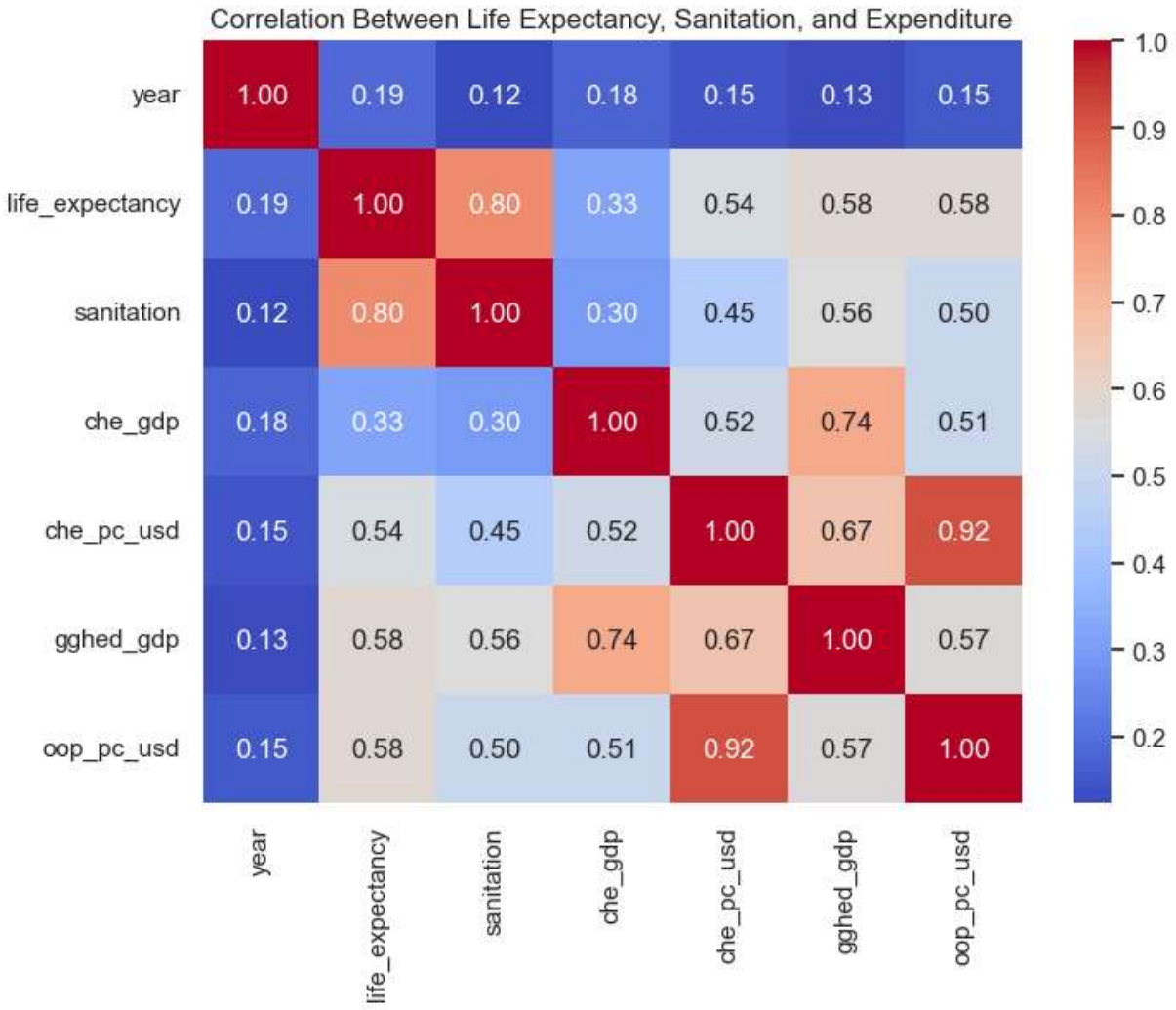
```
plt.figure(figsize=(8,5))
sns.scatterplot(data=clean_df, x='sanitation', y='life_expectancy', alpha=0.6, colo
plt.title('Life Expectancy vs Sanitation Coverage')
plt.xlabel('Sanitation Coverage (%)')
plt.ylabel('Life Expectancy (Years)')
plt.show()
```



correlation heatmap

In [104…

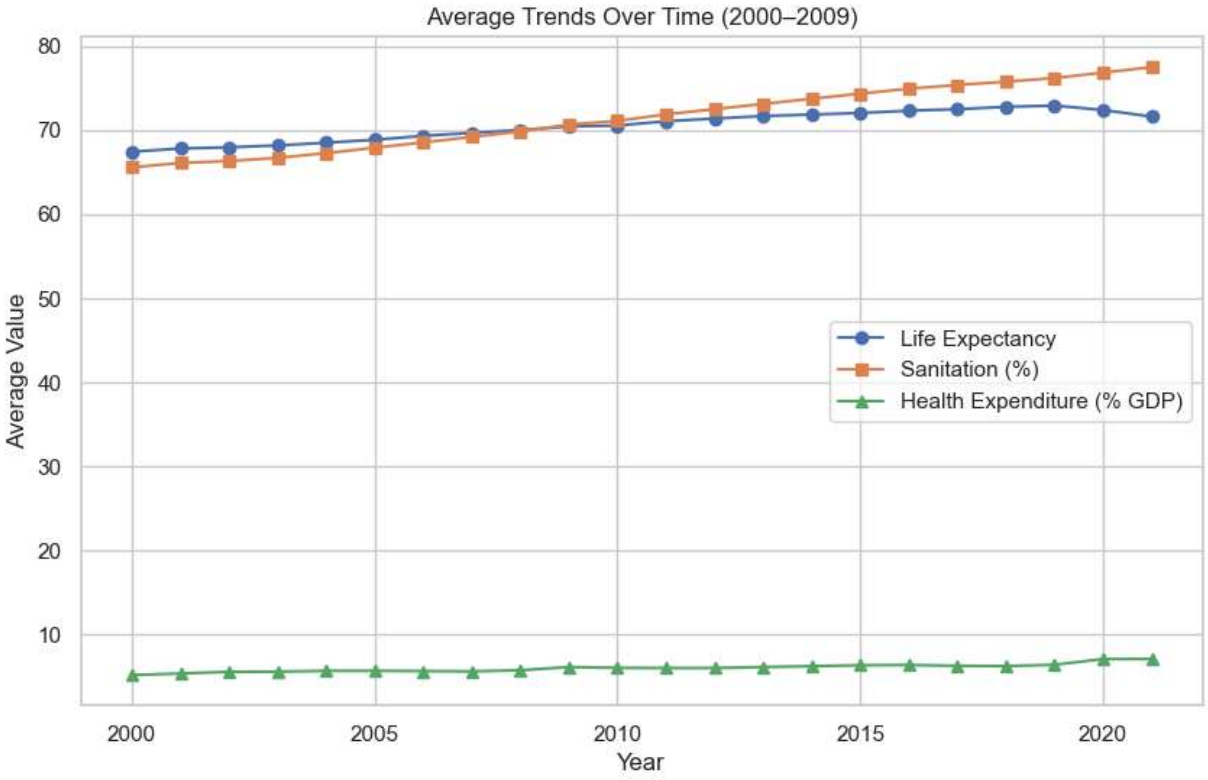
```
plt.figure(figsize=(8,6))
corr = clean_df.corr(numeric_only=True)
sns.heatmap(corr, cmap='coolwarm', annot=True, fmt=".2f")
plt.title('Correlation Between Life Expectancy, Sanitation, and Expenditure')
plt.show()
```



Trend over time

In [105…

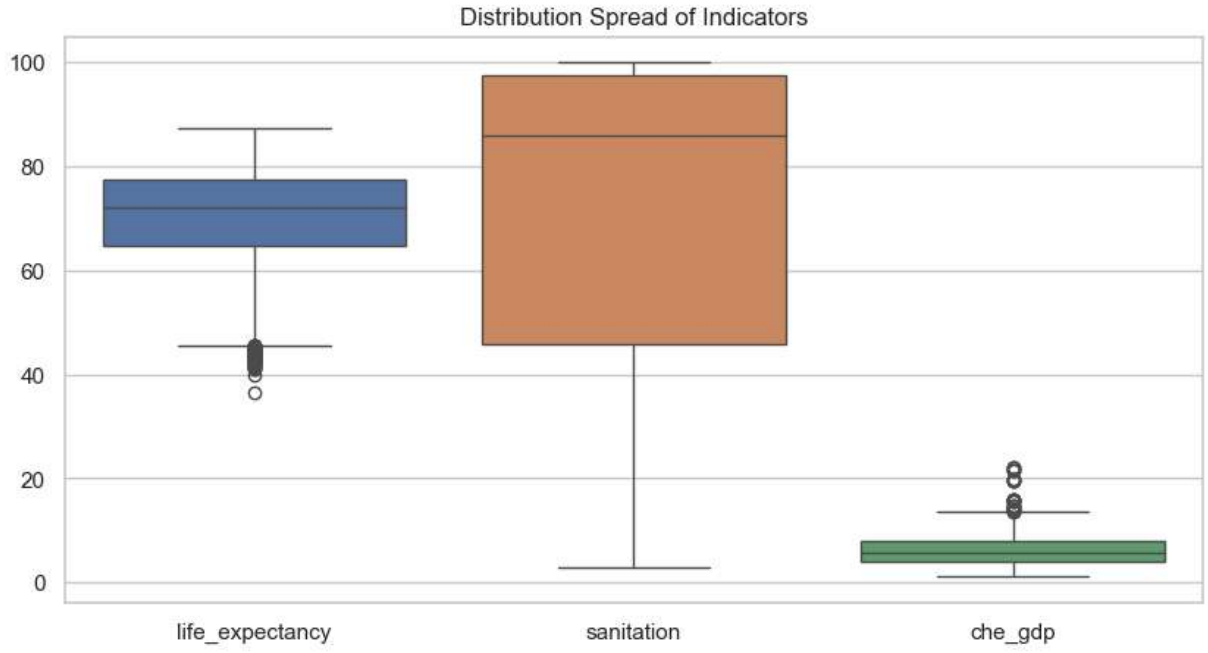
```
trend = clean_df.groupby('year')[['life_expectancy', 'sanitation', 'che_gdp']].mean

plt.figure(figsize=(10,6))
plt.plot(trend['year'], trend['life_expectancy'], marker='o', label='Life Expectanc
plt.plot(trend['year'], trend['sanitation'], marker='s', label='Sanitation (%)')
plt.plot(trend['year'], trend['che_gdp'], marker='^', label='Health Expenditure (%
plt.title('Average Trends Over Time (2000–2009)')
plt.xlabel('Year')
plt.ylabel('Average Value')
plt.legend()
plt.show()
```



Box plot

In [106…

```
plt.figure(figsize=(10,5))
sns.boxplot(data=clean_df[['life_expectancy', 'sanitation', 'che_gdp']])
plt.title('Distribution Spread of Indicators')
plt.show()
```



In [ ]: