Exploring the Relationship Between Life Expectancy, GDP per Capita, and Current Health Expenditure per Capita in 2020

In [27]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from scipy.stats import pearsonr
import statsmodels.api as sm
from statsmodels.formula.api import ols
import plotly.express as px
from scipy.stats import pearsonr
import plotly.graph_objects as go
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
```

In [2]:

```
df = pd.read_csv('Project.csv', encoding='ISO-8859-1')
```

In [3]:

```
df.head()
```

Out[3]:

	country	life_expectancy	healthcare_expenditure	gdp	che_per_capita
0	Afghanistan	62.6	0.1553	516.87	80.27
1	Albania	77.0	0.0680	5278.22	358.92
2	Algeria	74.5	0.0632	3354.15	211.98
3	Andorra	79.0	0.0905	37207.18	3367.25
4	Angola	62.3	0.0291	1639.95	47.72

In [4]:

```
df.dtypes
```

Out[4]:

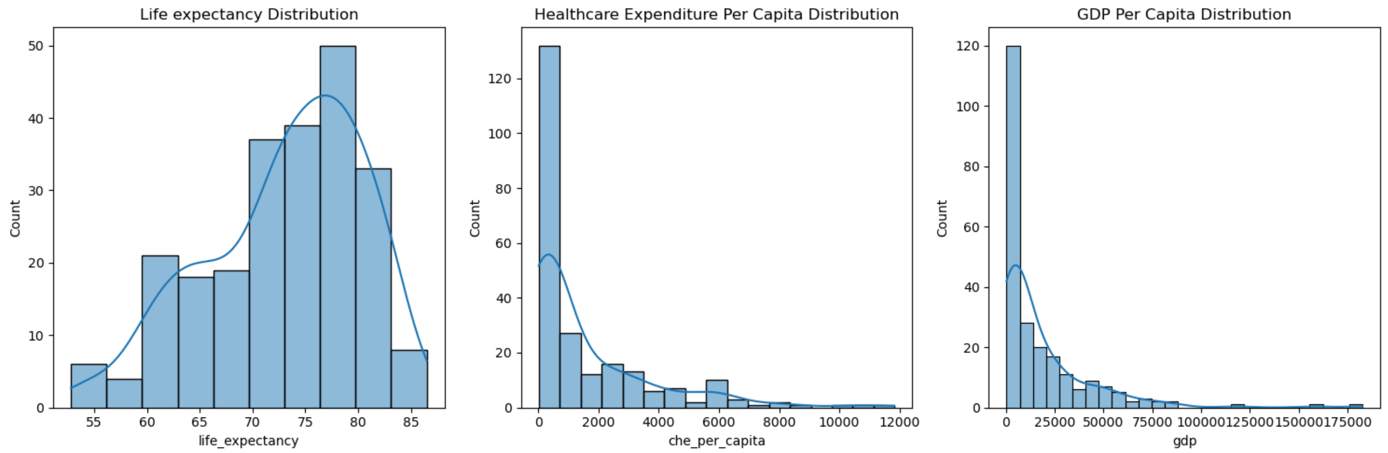
```
country                    object
life_expectancy           float64
healthcare_expenditure    float64
gdp                       float64
che_per_capita            float64
dtype: object
```

Distributions and Boxplots of GDP , Life Expectancy , and Healthcare Expenditure

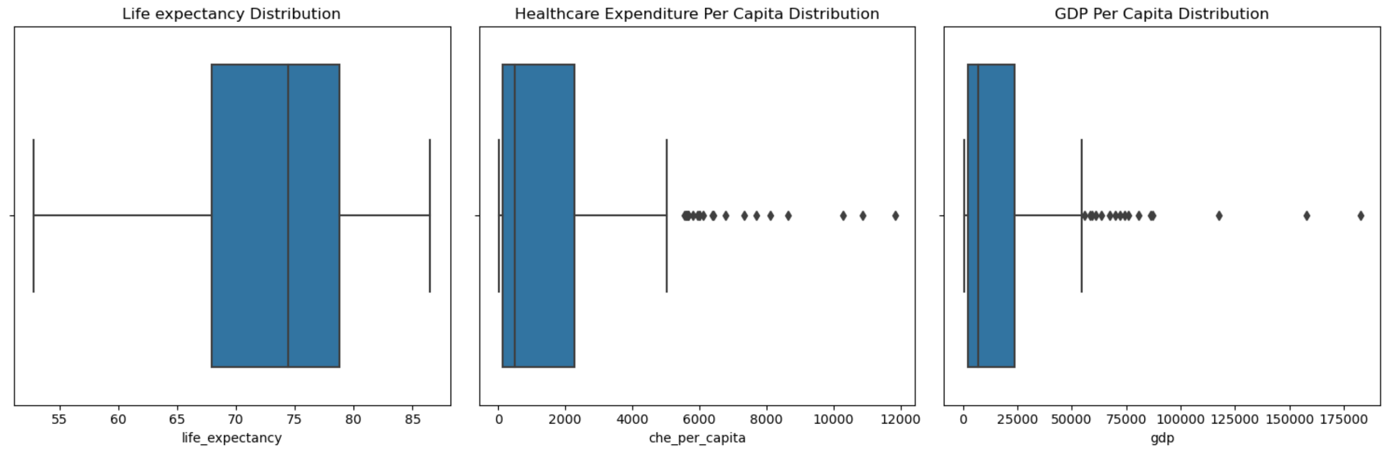
In [5]:

```
#
columns=['life_expectancy','che_per_capita','gdp']
titles=['Life expectancy Distribution','Healthcare Expenditure Per Capita Distribution',
fig,axes=plt.subplots(1, 3, figsize=(15, 5))

for col,title,ax in zip(columns,titles,axes):
    sns.histplot(df[col],kde=True,ax=ax)
    ax.set_title(title)
```

```
plt.tight_layout()
plt.show()
```



In [6]:
```
fig,axes=plt.subplots(1, 3, figsize=(15, 5))
for col,title,ax in zip(columns,titles,axes):
    sns.boxplot(x=df[col],ax=ax)
    ax.set_title(title)

plt.tight_layout()
plt.show()
```



Correlation between the variables

In [7]:
```
corr_matrix=df.corr()
corr_matrix
```

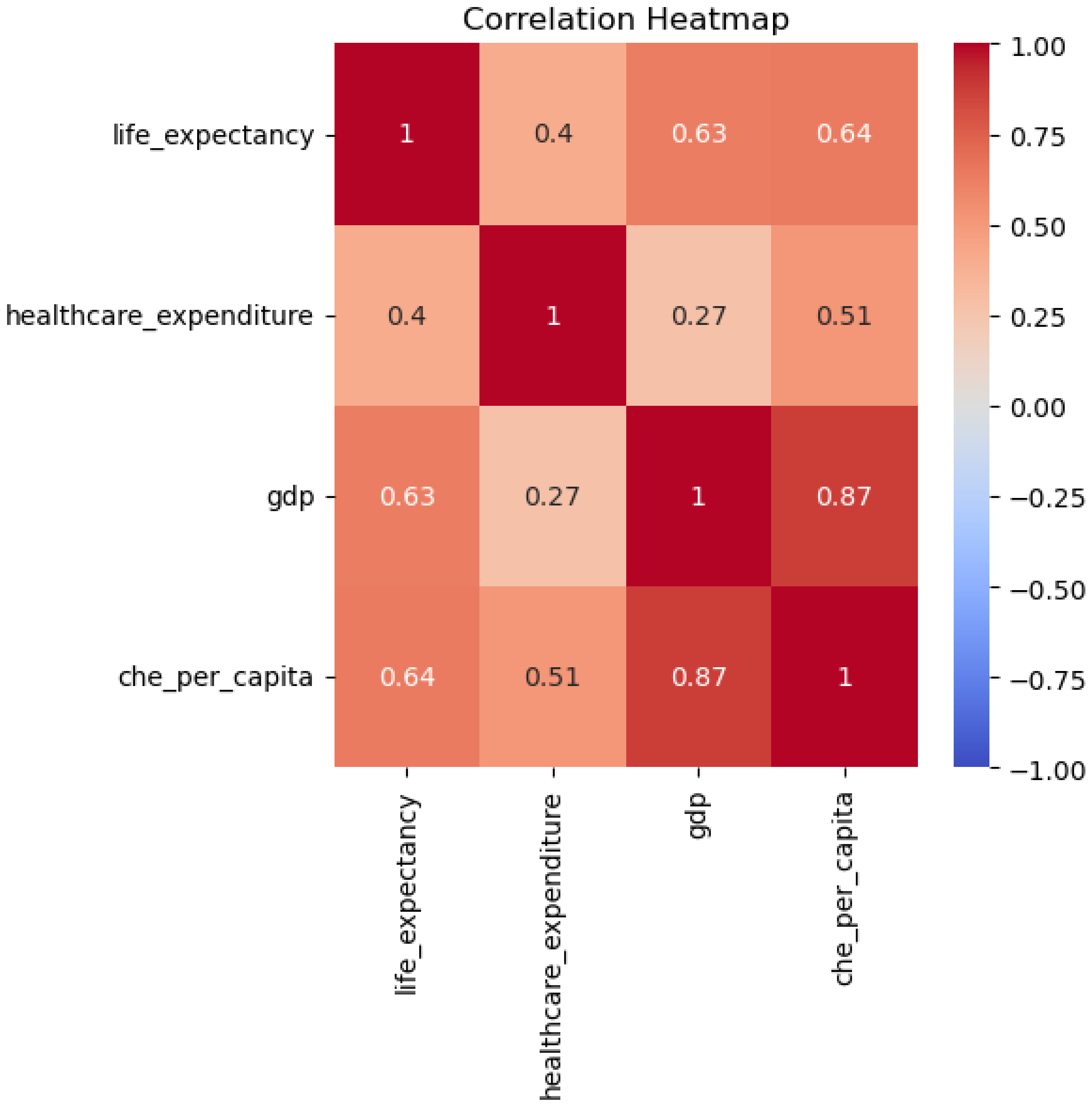
```
C:\Users\ttgmo\AppData\Local\Temp\ipykernel_19584\363827394.py:1: FutureWarning: The def
ault value of numeric_only in DataFrame.corr is deprecated. In a future version, it will
default to False. Select only valid columns or specify the value of numeric_only to sile
nce this warning.
  corr_matrix=df.corr()
```

Out[7]:

	life_expectancy	healthcare_expenditure	gdp	che_per_capita
life_expectancy	1.000000	0.399317	0.629102	0.635013
healthcare_expenditure	0.399317	1.000000	0.269825	0.510413
gdp	0.629102	0.269825	1.000000	0.869395
che_per_capita	0.635013	0.510413	0.869395	1.000000

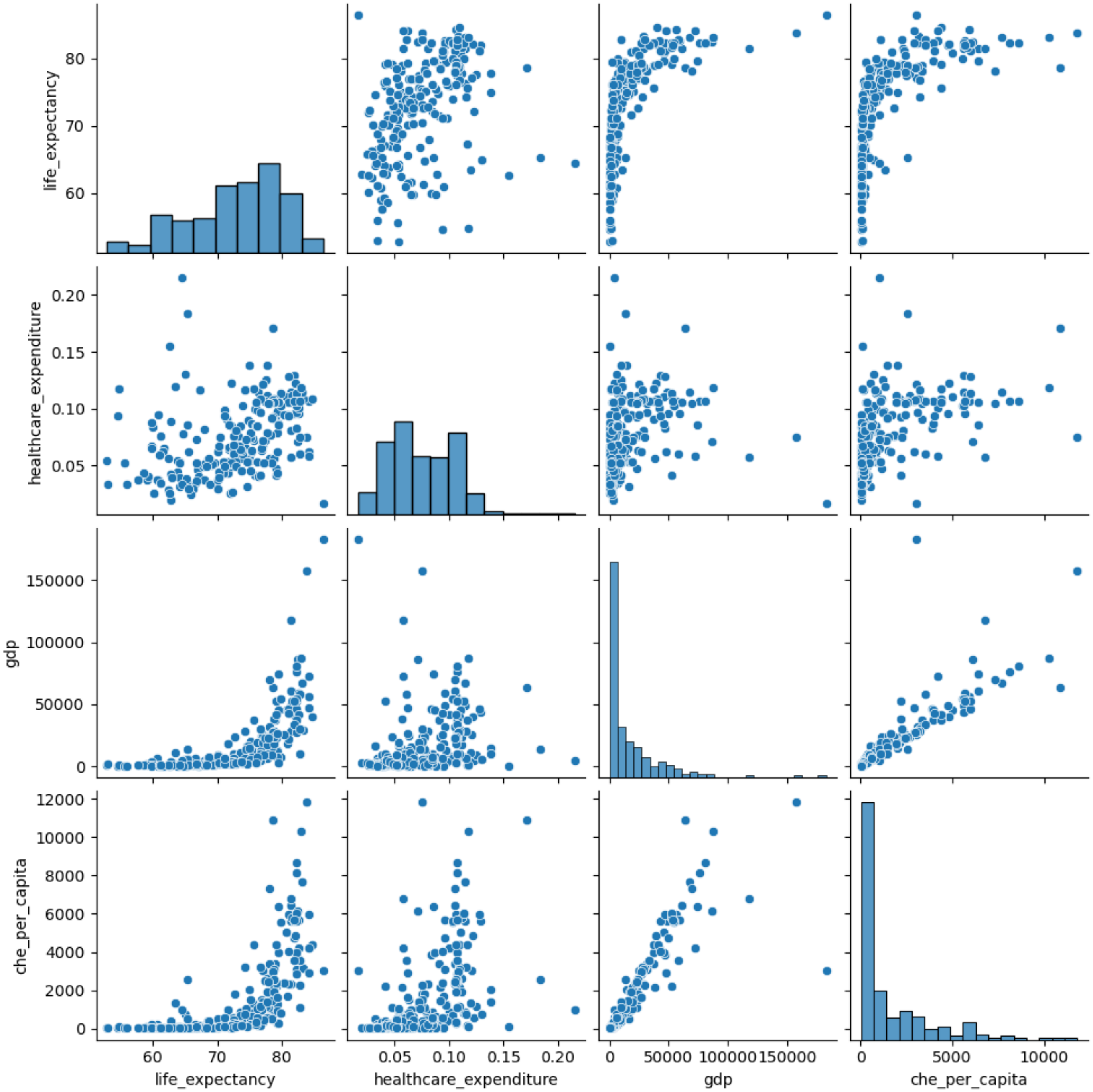
In [8]:
```
plt.figure(figsize=(5, 5))
```

```
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Correlation Heatmap')
plt.show()
```



In [9]:
```
sns.pairplot(df)
```

Out[9]:
```
<seaborn.axisgrid.PairGrid at 0x1a80e339330>
```



In [10]:

```python
import pandas as pd
import numpy as np
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Load your dataset (assuming it's in a CSV file)

# Select the independent variables
X = df[['gdp', 'che_per_capita']]

# Add a constant term to the independent variables
X = np.column_stack((np.ones(len(X)), X))

# Calculate the VIF for each independent variable
vif = pd.DataFrame()
vif['Variable'] = ['Intercept', 'gdp', 'che_per_capita']
vif['VIF'] = [variance_inflation_factor(X, i) for i in range(X.shape[1])]

# Print the VIF values
print(vif)
```

```
         Variable        VIF
```

```
0        Intercept  1.540358
1              gdp  4.095801
2   che_per_capita  4.095801
```

Bar Chart of incomes as bins

In [11]:

```python
## Creating bins for GPD Per Capita
bin_ranges = [0, 1045, 4095, 12695, float('inf')]
bin_labels = ['Low-income', 'Lower-middle-income', 'Upper-middle-income', 'High-income']
df['Income_Group'] = pd.cut(df['gdp'], bins=bin_ranges, labels=bin_labels)
df
```

Out[11]:

	country	life_expectancy	healthcare_expenditure	gdp	che_per_capita	Income_Group
0	Afghanistan	62.6	0.1553	516.87	80.27	Low-income
1	Albania	77.0	0.0680	5278.22	358.92	Upper-middle-income
2	Algeria	74.5	0.0632	3354.15	211.98	Lower-middle-income
3	Andorra	79.0	0.0905	37207.18	3367.25	High-income
4	Angola	62.3	0.0291	1639.95	47.72	Lower-middle-income
...	...	...	...	...	...	...
230	Wallis and Futuna	76.1	0.1060	3200.00	339.20	Lower-middle-income
231	Western Sahara	71.1	0.0490	1800.00	88.20	Lower-middle-income
232	Tanzania	64.1	0.0520	1010.00	52.52	Low-income
233	Bolivia	71.1	0.0940	6310.00	593.14	Upper-middle-income
234	Somalia	55.9	0.0340	580.00	19.72	Low-income

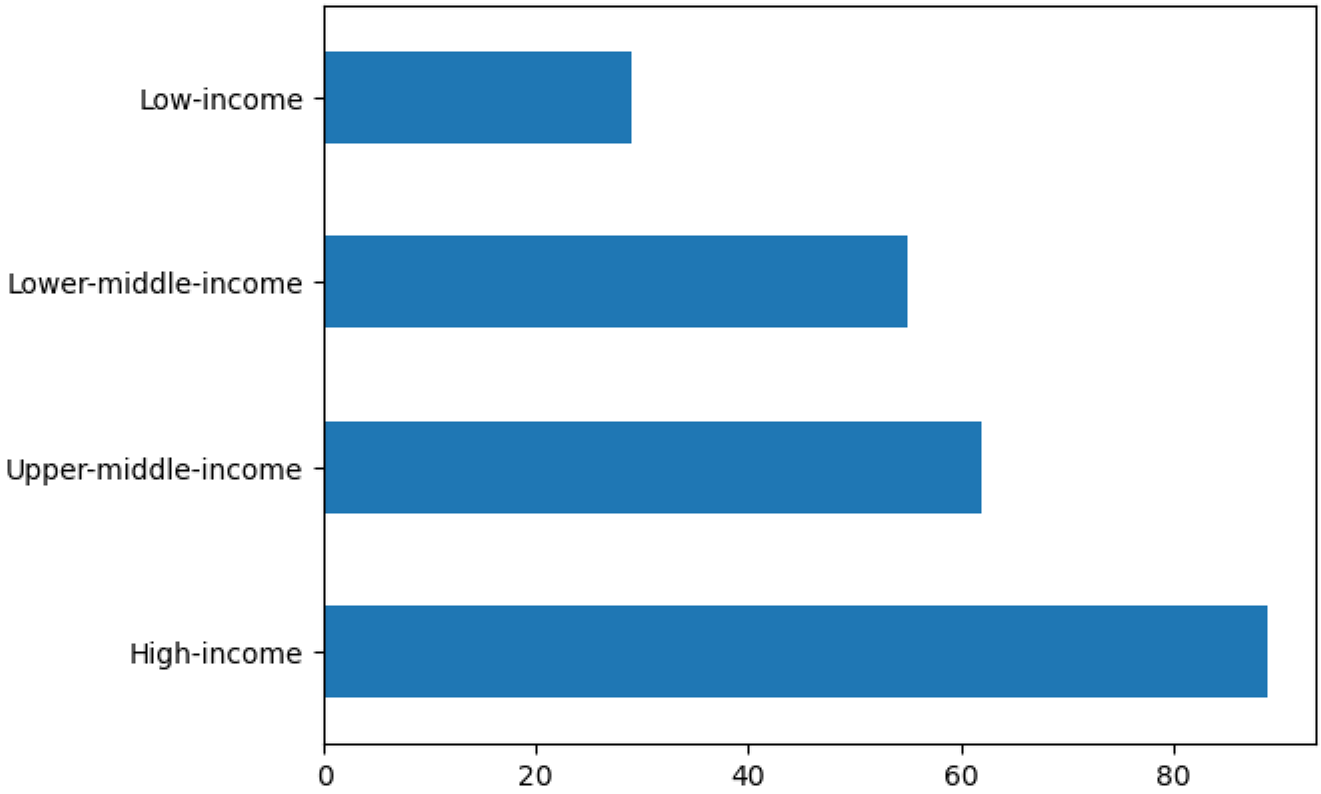
235 rows × 6 columns

In [12]:

```python
#Bar Chart of incomes as frequencies
df['Income_Group'].value_counts().plot(kind='barh')
```

Out[12]:

```
<Axes: >
```



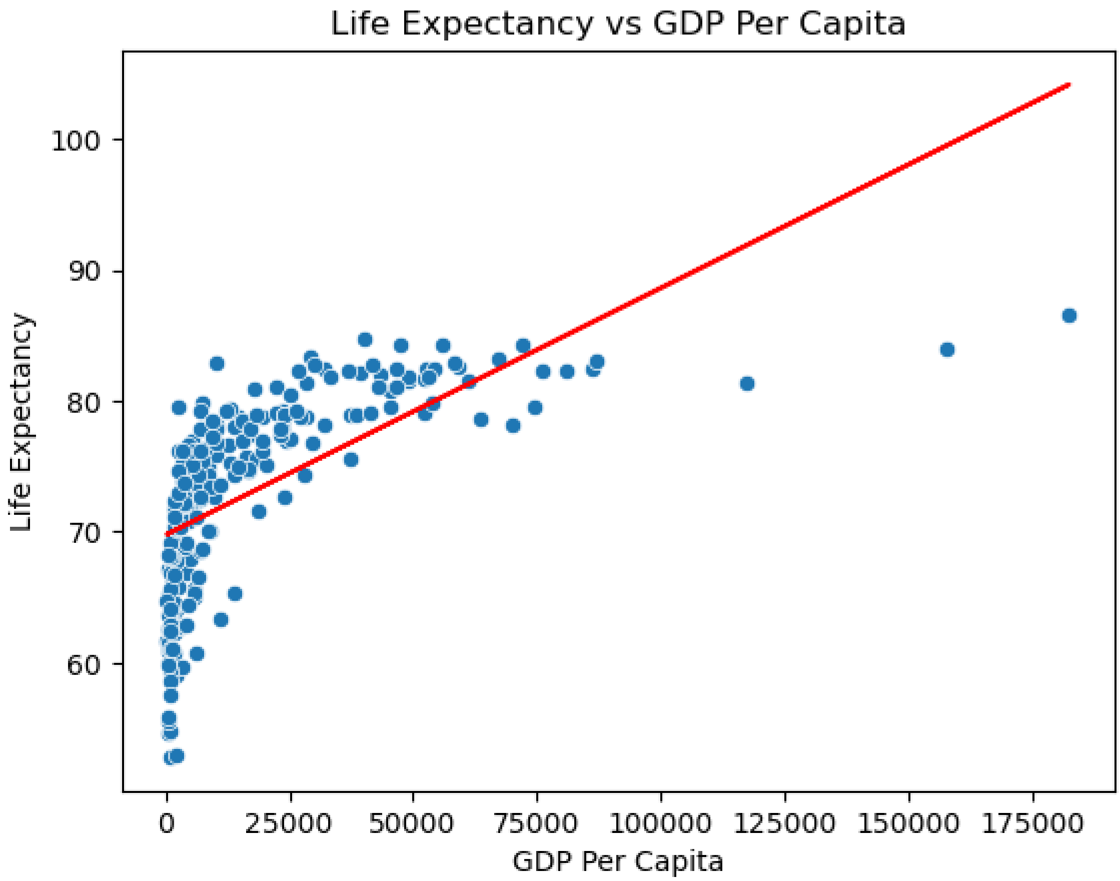
How does life expectancy vary across countries with different income levels?

In [13]:

```
# How does life expectancy vary across countries with different income levels?
sns.scatterplot(data=df, x='gdp', y='life_expectancy')
plt.title('Life Expectancy vs GDP Per Capita')
plt.xlabel('GDP Per Capita')
plt.ylabel('Life Expectancy')

# Calculate regression
reg = sm.OLS(df['life_expectancy'], sm.add_constant(df['gdp'])).fit()

# Plot regression line
plt.plot(df['gdp'], reg.params[0] + reg.params[1]*df['gdp'], 'r-')
```

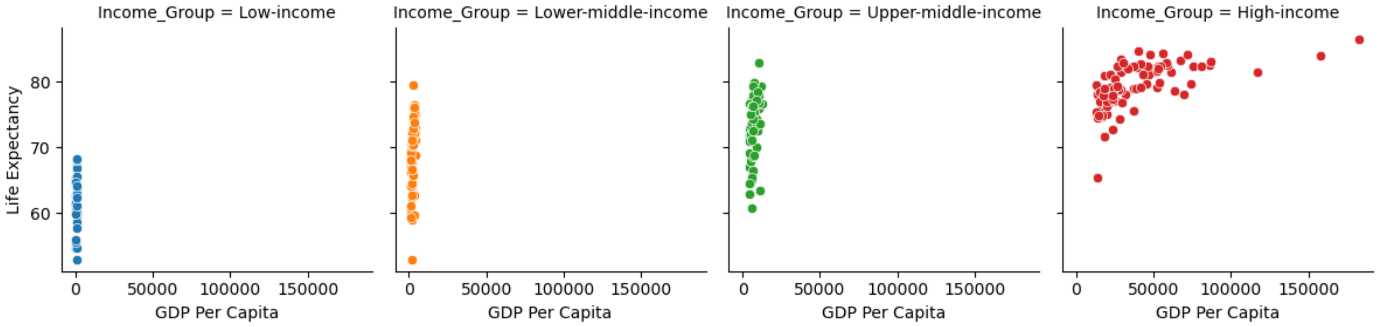
Out[13]:

```
[<matplotlib.lines.Line2D at 0x1a80f74cb50>]
```

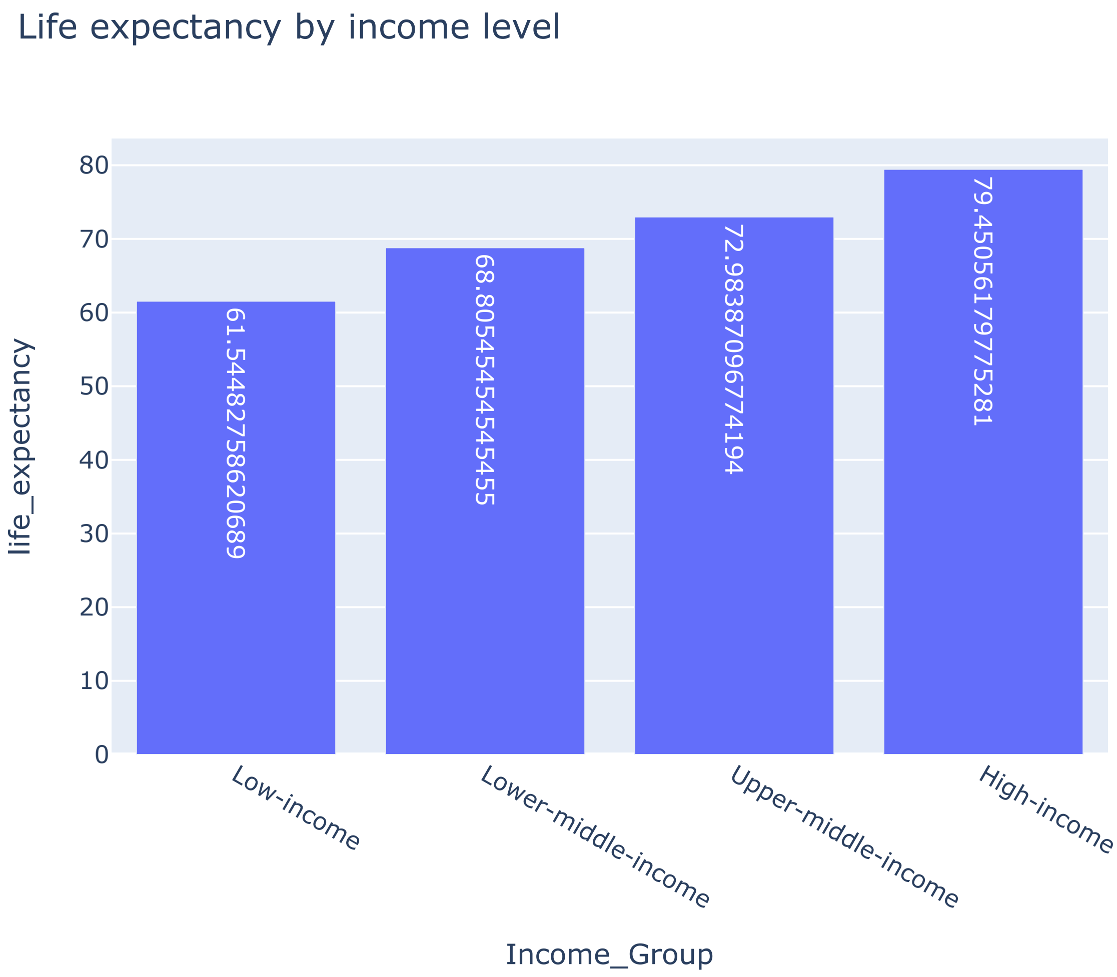


```
In [14]: #FacetGrid of the Income
         g = sns.FacetGrid(df, col='Income_Group',hue='Income_Group')
         g.map(sns.scatterplot, 'gdp', 'life_expectancy')
         g.set_axis_labels('GDP Per Capita', 'Life Expectancy')
```

```
Out[14]: <seaborn.axisgrid.FacetGrid at 0x1a80f74d7b0>
```



```
In [15]: # Group by Income_Group
         low_income = df[df['Income_Group'] == 'Low-income']
         lower_middle_income = df[df['Income_Group'] == 'Lower-middle-income']
         upper_middle_income = df[df['Income_Group'] == 'Upper-middle-income']
         high_income = df[df['Income_Group'] == 'High-income']

         life_exp = df.groupby('Income_Group')['life_expectancy'].mean()
         fig = px.bar(life_exp,
                      x=life_exp.index,
                      y='life_expectancy',
                      text='life_expectancy',
                      title='Life expectancy by income level')

         # Show plot
         fig.show()
```



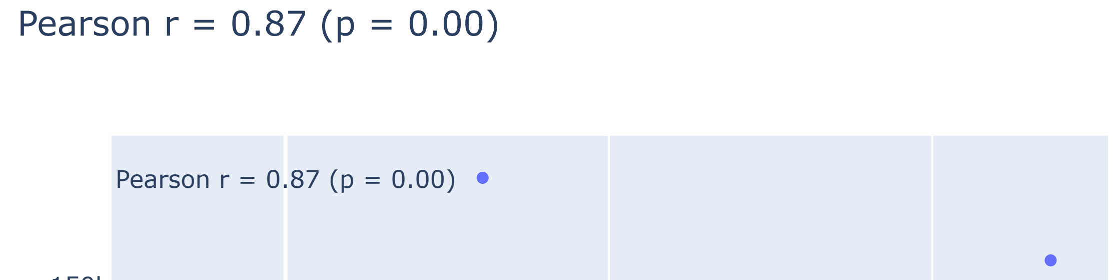
How limited resources and healthcare access in low-income countries may influence life expectancy?

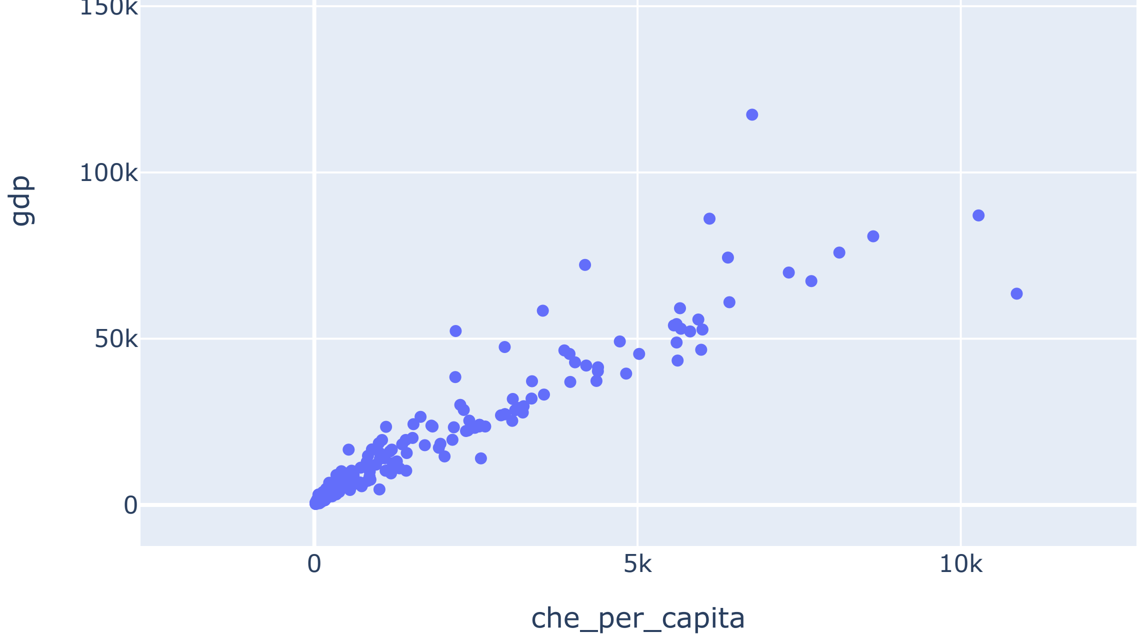
- In low-income countries, life expectancy suffers due to limited resources, inadequate healthcare, and prevalent violence. By concentrating on targeted investments in four key areas—healthcare, nutrition, water/sanitation, and peace—we can bring about transformative change. These strategic efforts can improve the living conditions and prospects for millions, resulting in longer, healthier lives, and fostering a more just and prosperous world for everyone.

What is the relationship between GDP per capita and current health expenditure per capita ?

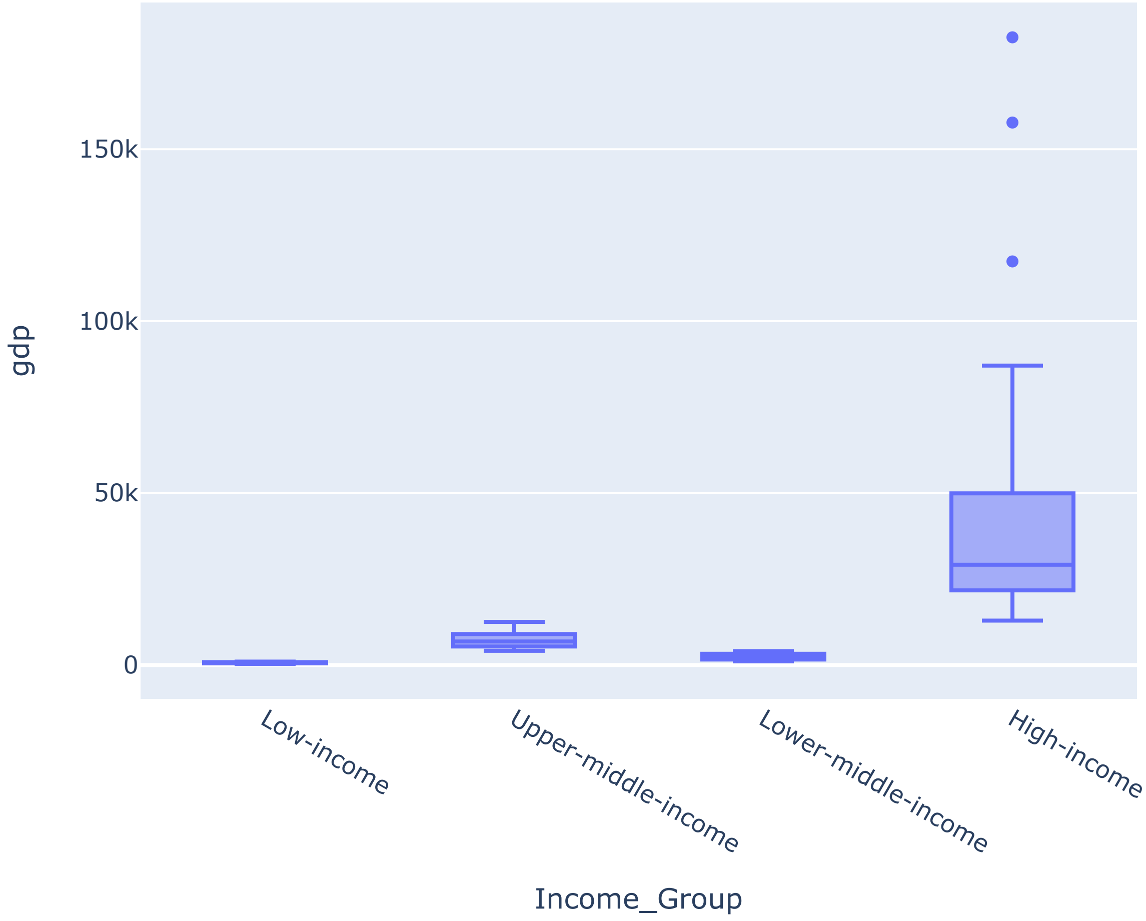
Before removing the outliers

```
In [16]: r, p = pearsonr(df['che_per_capita'], df['gdp'])
         fig = px.scatter(df, x='che_per_capita', y='gdp', title=f'Pearson r = {r:.2f} (p = {p:.2
         fig.add_annotation(x=df['che_per_capita'].min(), y=df['gdp'].max(),
                            text=f'Pearson r = {r:.2f} (p = {p:.2f})',
                            showarrow=False)
```



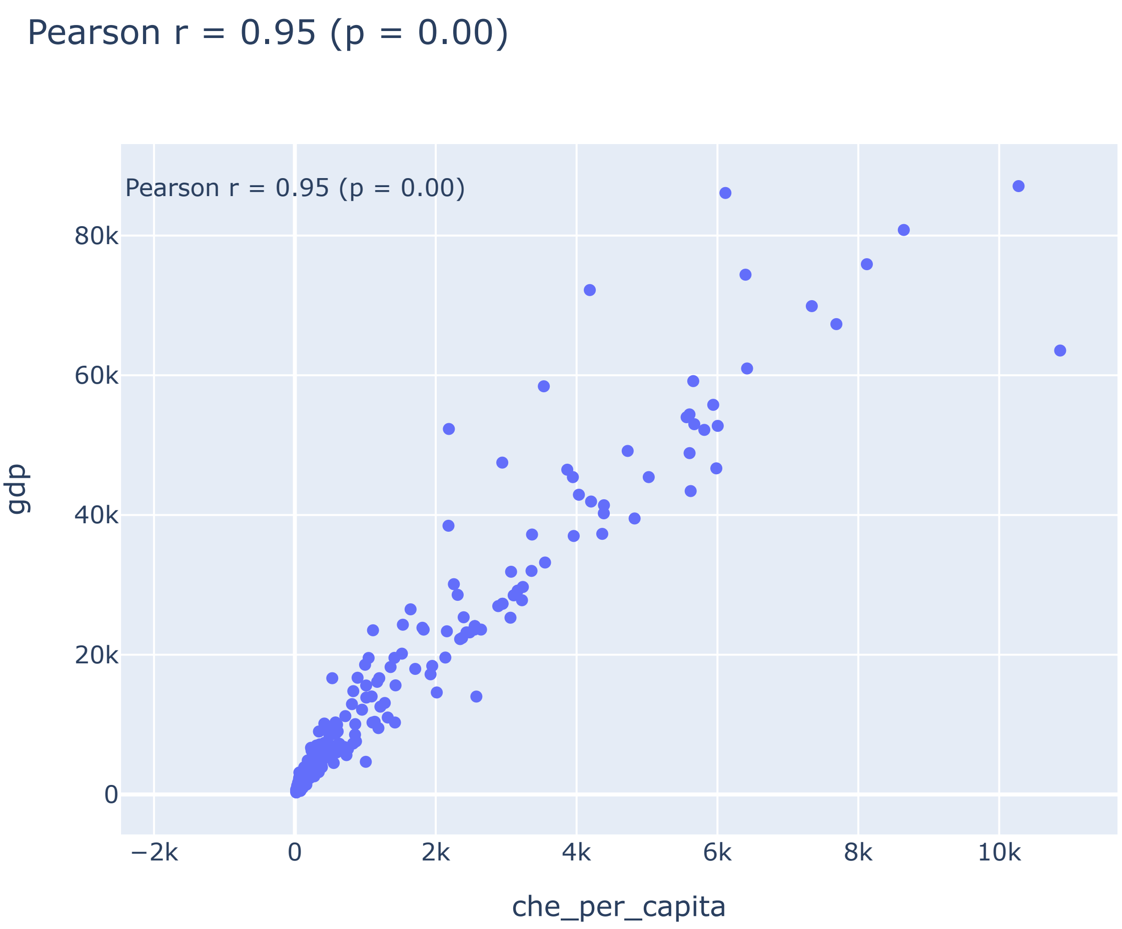


```
In [17]: # Identifying the outliers based on GDP
         px.box(df,x='Income_Group',y='gdp')
```



After removing outliers

```
In [18]: df_outlier = df.drop(df[df['country'].isin(['Monaco', 'Luxembourg','Liechtenstein'])].in
         r, p = pearsonr(df_outlier['che_per_capita'], df_outlier['gdp'])
         fig = px.scatter(df_outlier, x='che_per_capita', y='gdp', title=f'Pearson r = {r:.2f} (p
         fig.add_annotation(x=df_outlier['che_per_capita'].min(), y=df_outlier['gdp'].max(),
                            text=f'Pearson r = {r:.2f} (p = {p:.2f})',
                            showarrow=False)
```



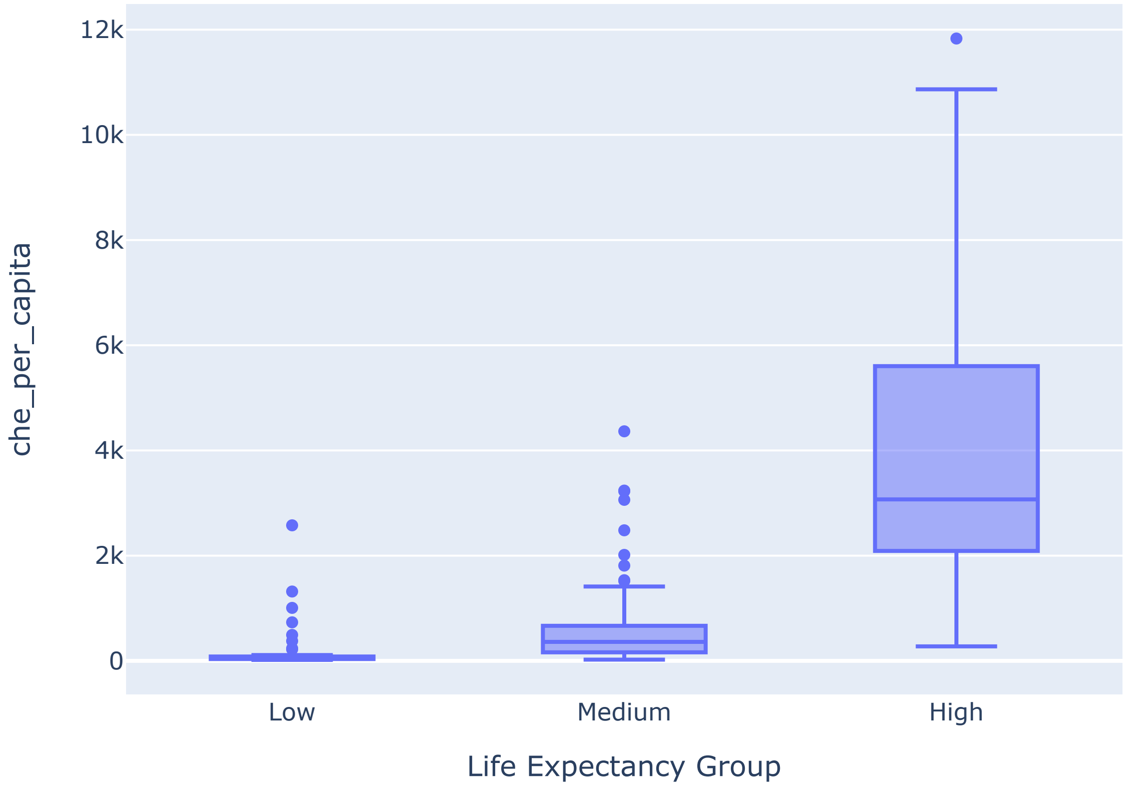
There is a positive relationship between GDP per capita and current health expenditure per capita. This means that wealthier nations tend to spend more on healthcare than poorer nations. This is likely due to a number of factors, including: -Wealthier nations have more resources to invest in healthcare, such as money, personnel, and technology. -Wealthier nations have higher rates of chronic diseases, such as heart disease, cancer, and diabetes, which require more expensive treatment. -Wealthier nations have higher expectations for healthcare, and are willing to pay more for it.

How does the distribution of current health expenditure per capita differ between countries with low, medium, and high life expectancy?

In [19]:

```
#Using boxplots to observe where are the outliers segmented in which groups
labels=['Low','Medium','High']
life_expectancy_bins=[0,65.55,77.64,100]
df['Life Expectancy Group']=pd.cut(df.life_expectancy,bins=life_expectancy_bins,labels=l
fig=px.box(df,x='Life Expectancy Group',y='che_per_capita',title='Life expectancy by Cur
fig.show()
```

Life expectancy by Current Health Expenditure per Capita



To translate greater healthcare expenditure into meaningful life expectancy gains, countries must take an integrated long-term approach across all aspects of public health. Simply increasing budgets will not achieve outcomes; money must be well-spent and targeted as part of a multi-pronged strategy to make a difference. This includes ensuring equitable access to care for all through universal healthcare, reduced costs for the disadvantaged, and fair resource distribution so new funds actually reach those in need. Healthcare systems must also incentive efficiency and accountability, reducing waste and unnecessary administration, limiting corruption, and rewarding good outcomes and preventive care. Performance metrics and oversight maximize the impact of new resources. Public-private hybrid systems often strike the right balance. Further, healthcare funding must be part of a broader public health strategy including investments in education, sanitation, poverty reduction, health education, and environment, which are equally essential to well-being. While more money for treatments matters, health starts before illness, and prevention has higher returns. Countries need an integrated long-term approach recognizing people live beyond clinics alone. Strong public health practices, social programs, and gradual funding increases over time as economies grow create lasting success, while quick fixes lead to disappointment. Overall prosperity and robust public health strategies drive outcomes, not budgets and technologies alone. Well-spent money, not just more of it, is key. Increased healthcare expenditure can substantially improve life expectancy only if paired with wider reforms enabling new resources to be strategically invested over generations. Access for all, efficiency, accountability, education, poverty reduction, and environment investments alongside medical funding over time are required to realize the benefits of greater investment in a sustainable way.

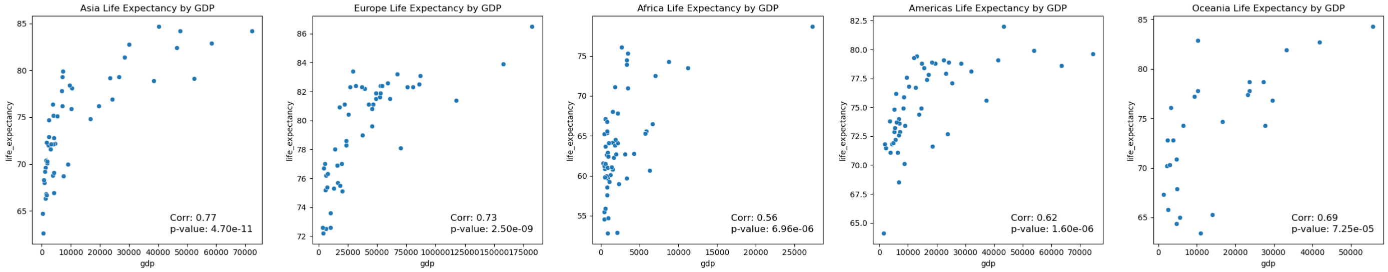
What is the relationship between life expectancy and economic status within different world regions?

In [20]:
```
df_regions=pd.read_csv('Reigons.csv')
```

In [21]:
```
merged_df = df.merge(df_regions[['Country', 'Region', 'Subregion']], left_on='country',
merged_df.drop('Country', axis=1, inplace=True)
```

In [22]:

```
Regions = ['Asia', 'Europe', 'Africa', 'Americas', 'Oceania']
Titles = ['Asia Life Expectancy by GDP', 'Europe Life Expectancy by GDP', 'Africa Life E
fig, axes = plt.subplots(1, 5, figsize=(25, 5))

for reg, title, ax in zip(Regions, Titles, axes):
    region_df = merged_df[merged_df['Region'] == reg]
    sns.scatterplot(data=region_df, x='gdp', y='life_expectancy', ax=ax)
    ax.set_title(title)
    correlation, p_value = pearsonr(region_df['gdp'], region_df['life_expectancy'])
    ax.text(0.6, 0.1, f'Corr: {correlation:.2f}', transform=ax.transAxes, fontsize=12)
    ax.text(0.6, 0.05, f'p-value: {p_value:.2e}', transform=ax.transAxes, fontsize=12)

plt.tight_layout()
plt.show()
```



The relationship between life expectancy and wealth (GDP per capita) varies across different world regions. Asia, Europe, and Oceania exhibit a moderately strong to strong positive correlation, indicating that increased wealth tends to have a noticeable impact on life expectancy. This suggests that investments in healthcare, infrastructure, and other factors contributing to overall well-being have a significant influence on the life expectancy in these regions. However, Africa and the Americas show a moderate positive correlation, implying that although there is a relationship between wealth and life expectancy, other factors might be playing a more prominent role in these regions.

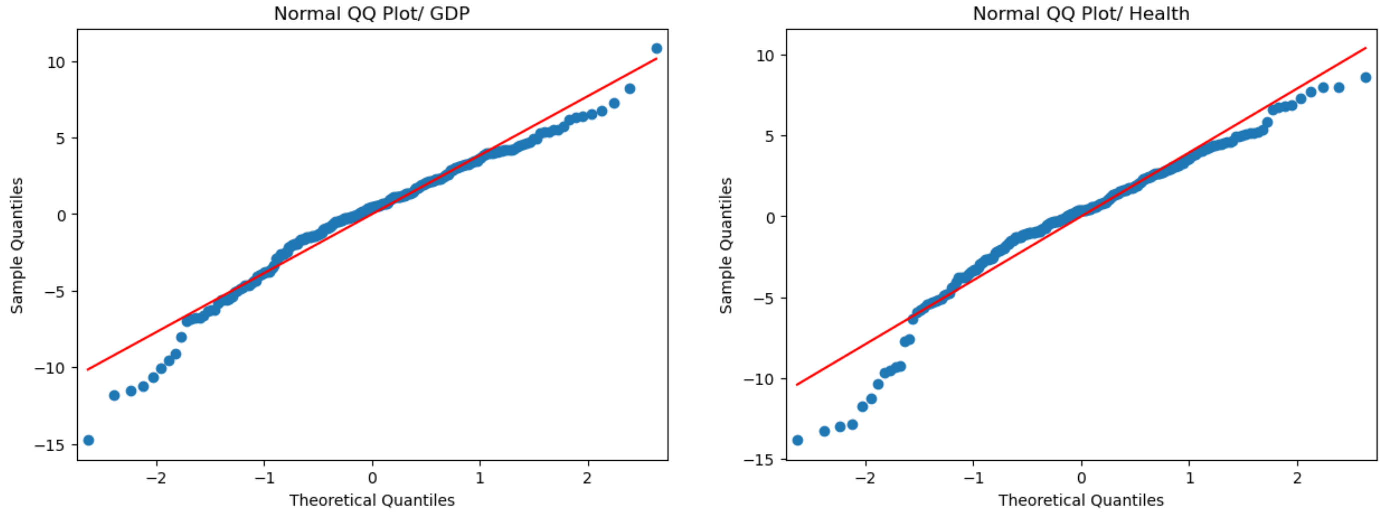
Several factors can strengthen or weaken the relationship between life expectancy and wealth within regions, including poverty, inequality, infrastructure, and disease burden. High levels of poverty and income inequality can limit the impact of wealth on life expectancy, as resources might not be evenly distributed across the population or effectively utilized to improve living standards. Additionally, inadequate infrastructure, such as healthcare facilities, sanitation systems, and transportation networks, can hinder the benefits of increased wealth on life expectancy. Finally, regions with a high disease burden might experience a weaker relationship between life expectancy and wealth, as addressing these health problems may require substantial resources and time. In summary, while the correlation coefficients indicate a positive relationship between life expectancy and wealth in all regions, the strength of this relationship is influenced by various factors that affect the overall well-being of the population.

Can we predict life expectancy based on GDP per capita and current health expenditure per capita?

In [28]:

```
merged_df['log_gdp'] = np.log(merged_df['gdp'])
merged_df['log_che_per_capita'] = np.log(merged_df['che_per_capita'])
fig,axes=plt.subplots(1, 2, figsize=(15, 5))
GDP=ols(formula='life_expectancy ~ log_gdp',data=merged_df).fit()
Health=ols(formula='life_expectancy ~ log_che_per_capita',data=merged_df).fit()
gdp_resid=GDP.resid
health_resid=Health.resid
List = [gdp_resid, health_resid]
titles = ['Normal QQ Plot/ GDP', 'Normal QQ Plot/ Health']
```

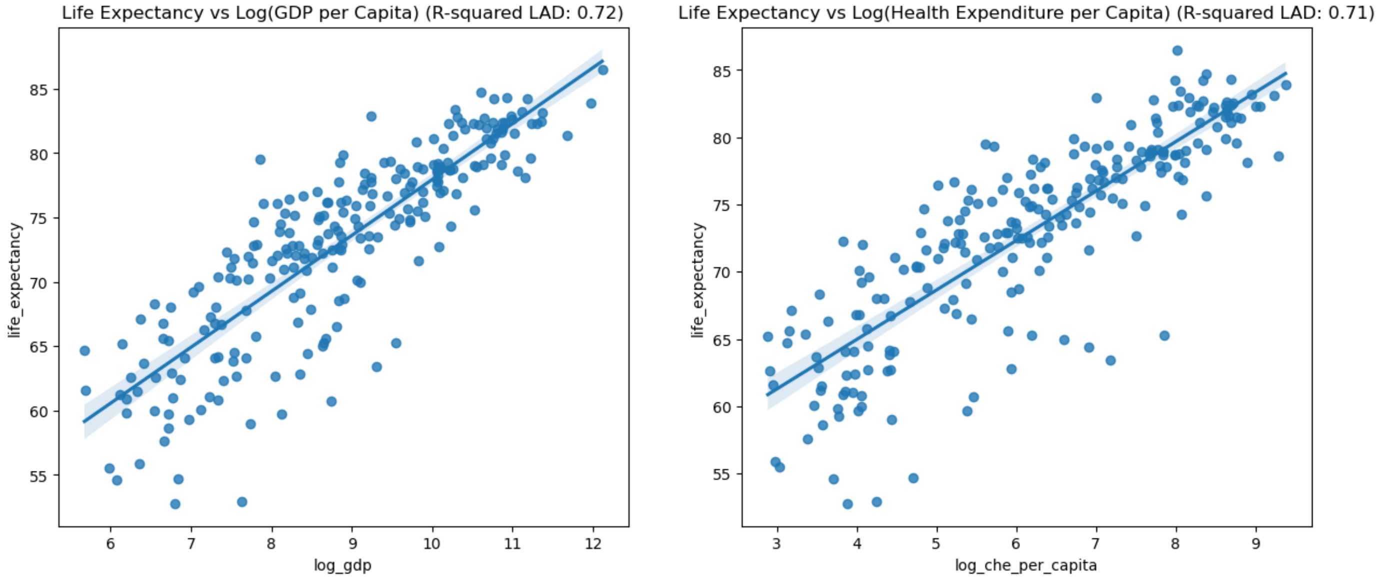
```
for residuals, ax, title in zip(List, axes, titles):
    sm.qqplot(residuals, line='s', ax=ax)
    ax.set_title(title)
```



The presence of outliers in the qq plot indicates that the residuals are not normally distributed, which is a violation of one of the assumptions of simple linear regression. This may lead to inaccurate results, so it is important to address the issue of outliers before proceeding with the analysis.

In [29]:

```
# Your data processing code here (merged_df creation)

merged_df['log_gdp'] = np.log(merged_df['gdp'])
merged_df['log_che_per_capita'] = np.log(merged_df['che_per_capita'])

data_vars = [('log_gdp', 'Life Expectancy vs Log(GDP per Capita)'),
             ('log_che_per_capita', 'Life Expectancy vs Log(Health Expenditure per Capit

fig, axes = plt.subplots(1, 2, figsize=(15, 6))

for ax, (var, title) in zip(axes, data_vars):
    x = merged_df[var].values.reshape(-1,1)
    y = merged_df['life_expectancy'].values

    # LAD Regression (Quantile Regression with quantile = 0.5)
    lad_model = smf.quantreg(f'life_expectancy ~ {var}', data=merged_df)
    lad_result = lad_model.fit(q=0.5)
    y_pred = lad_result.predict(merged_df[var])

    # Calculate R-squared
    ss_res = np.sum((y - y_pred) ** 2)
    ss_tot = np.sum((y - np.mean(y)) ** 2)
    r_squared = 1 - (ss_res / ss_tot)

    sns.regplot(x=merged_df[var], y=merged_df['life_expectancy'], ax=ax)
    ax.set_title(f'{title} (R-squared LAD: {r_squared:.2f})')


plt.show()
```



In this study, I investigated the relationship between GDP per capita, health expenditure per capita, and life expectancy. I used a robust regression method, specifically Least Absolute Deviations (LAD) Regression / Quantile Regression, to address the issue non-normality in the data as well as MLR(Multiple Linear Regression) couldn't be used as there multicollinearity between gdp per capita and health expenditure per capita.

My results showed that both GDP per capita and health expenditure per capita have a strong positive relationship with life expectancy. The R-squared values for LAD regression were 0.72 and 0.71, respectively. This suggests that these two variables can explain 72% and 71% of the variation in life expectancy, respectively.

While this analysis focused only on GDP per capita and health expenditure per capita, it is important to be aware that other factors may also influence life expectancy. There are other factors, such as education level, access to healthcare, and environmental conditions, that can also influence life expectancy.

Overall, my findings suggest that GDP per capita and health expenditure per capita are important factors in determining life expectancy. However, it is important to consider other factors as well when making predictions about life expectancy.

In [ ]:

```
merged_df.to_csv('Tableau.csv', index=False)
```