Case Study: World Happiness Report 2016 – Data Analysis & Visualization

This project is part of the IBM Data Analyst Professional Certificate. It involves data preparation, exploration, and visualization based on the World Happiness Report.

Project Goals or Objectives

These include:

- Analyze the happiness metrics of 2016 across regions.
- Discover how economic, social, and health factors influence happiness.
- Build interactive visualizations and a summary dashboard.

Main Tasks in the Project

These include:

1. There might be a few missing values in the dataset. Data cleaning will be a part of the assignment.
2. You have to perform exploratory data analysis to draw keen insights on the data:
   - Identify the GDP per capita and Healthy Life Expectancy of the top 10 countries.and represent it as a bar chart.
   - Find the correlation between the Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, and Happiness Score.
   - Create a scatter plot to identify the effect of GDP per Capita on Happiness Score in various Regions.
   - Create a pie chart to present Happiness Score by region.
   - Create a map to display GDP per capita of countries and include Healthy Life Expectancy as a tooltip.
3. Create a dashboard with at least four of the above visualizations.
4. Present insights, patterns, and observations. Write a short executive summary.

About the Dataset

This project uses data from the **World Happiness Report**, a widely cited global survey that ranks countries based on their citizens' perceived well-being. The report draws on recent research in the science of happiness to explain variations in life satisfaction across nations. The dataset is publicly available on **Kaggle** and is released under the **CC0: Public Domain license**, making it freely usable for analysis and visualization.

Dataset Attributes

Variable	Description
`Country`	Name of the country
`Region`	Region the country belongs to
`Happiness Rank`	Rank of the country based on the Happiness Score
`Happiness Score`	A metric measured in 2016 by asking people: "How would you rate your happiness?"
`Lower Confidence Interval`	Lower bound of the confidence interval for the Happiness Score
`Upper Confidence Interval`	Upper bound of the confidence interval for the Happiness Score
`Economy (GDP per Capita)`	The extent to which GDP contributes to the calculation of the Happiness Score
`Family`	The extent to which family contributes to the calculation of the Happiness Score
`Health (Life Expectancy)`	The extent to which life expectancy contributes to the calculation of the Happiness Score

Variable	Description
`Freedom`	The extent to which freedom contributes to the calculation of the Happiness Score
`Trust (Government Corruption)`	The extent to which trust in government contributes to the calculation of the Happiness Score
`Generosity`	The extent to which generosity contributes to the calculation of the Happiness Score
`Dystopia Residual`	Represents unexplained components of the score. Reflects how the six factors under/over the average value is approximately zero globally.

Tools Used

These include:

- Python (Pandas, NumPy, Matplotlib, Seaborn, Plotly).
- Jupyter Notebook.
- Streamlit or Plotly Dash (for dashboarding).
- GitHub for version control and project public.

Key Visualizations

These include:

- Top 10 Happiest Countries by GDP and Life Expectancy (Bar Chart).
- Correlation Heatmap of Factors Influencing Happiness.
- GDP vs Happiness by Region (Scatter Plot).
- Happiness Score by Region (Pie Chart).
- Interactive Global Map of GDP & Life Expectancy.

Summary & Key Takeaways

- Countries with high GDP, strong healthcare, and family support consistently score higher in happiness.
- Western Europe leads in both GDP and overall happiness.
- Regions with lower economic output tend to have lower happiness scores, though cultural and social support may also play a role.
- Correlation matrix confirms that economy, health, and family are the strongest influencers of happiness.

The interactive dashboard below provides a data-driven view into what makes people happy — and how different regions compare globally.

Key Insights

GDP per Capita, **Life Expectancy**, and **Family** are the top predictors of national happiness.

- **Western Europe** ranks highest in happiness, driven by strong economic and healthcare indicators.
- **Sub-Saharan Africa** shows the lowest scores overall, with lower values across multiple contributing factors.
- Regions with **more countries** (like Africa) contribute a larger share to global happiness totals even if per-country scores are lower.

This project demonstrates how data storytelling through visual analytics can highlight global well-being patterns and support evidence-based insights into what makes people happy.

IBM Certification

Please verify here

In [ ]:

In [3]:

```
import pandas as pd

# Read the locally uploaded CSV file
# df = pd.read_csv("World_Happiness_Report2016.csv")

import pandas as pd

# Load the dataset
url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMSkillsNetwork-AI0272EN-SkillsNetwork/labs/dataset/2016.csv"
df = pd.read_csv(url)

# Display the first 5 rows
df.head()
```

Out[3]:

	Country	Region	Happiness Rank	Happiness Score	Lower Confidence Interval	Upper Confidence Interval	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	(Go C
0	Denmark	Western Europe	1	7.526	7.460	7.592	1.44178	1.16374	0.79504	0.57941	
1	Switzerland	Western Europe	2	7.509	7.428	7.59	1.52733	1.14524	0.86303	0.58557	
2	Iceland	Western Europe	3	7.501	7.333	7.669	1.42666	1.18326	0.86733	0.56624	
3	Norway	Western Europe	4	7.498	7.421	7.575	1.57744	1.12690	0.79579	0.59609	
4	Finland	Western Europe	5	7.413	7.351	7.475	1.40598	1.13464	0.81091	0.57104	

In [5]:

```
# Check the structure and columns of the dataset
print("Shape of dataset:", df.shape)
print("\nColumn names:\n", df.columns.tolist())
```

```
Shape of dataset: (157, 13)

Column names:
 ['Country', 'Region', 'Happiness Rank', 'Happiness Score', 'Lower Confidence Interval', 'Upper Confidence Interval', 'Economy (GDP per Capita)', 'Family', 'Health (Life E
xpectancy)', 'Freedom', 'Trust (Government Corruption)', 'Generosity', 'Dystopia Residual']
```

In [7]:

```
# Check for missing/null values
missing_values = df.isnull().sum()
print("Missing values in each column:\n", missing_values)
```

```
Missing values in each column:
 Country                          0
Region                           0
Happiness Rank                   0
Happiness Score                  0
Lower Confidence Interval        4
Upper Confidence Interval        2
Economy (GDP per Capita)         1
Family                           0
Health (Life Expectancy)         2
Freedom                          0
Trust (Government Corruption)    0
Generosity                       0
Dystopia Residual                0
dtype: int64
```

In [9]:

```
# Check data types and summary statistics
print("\nData types:\n", df.dtypes)
print("\nSummary statistics:\n", df.describe())
```

```
Data types:
 Country                          object
Region                           object
Happiness Rank                    int64
Happiness Score                 float64
Lower Confidence Interval       float64
Upper Confidence Interval        object
Economy (GDP per Capita)         object
Family                          float64
Health (Life Expectancy)         object
Freedom                          object
Trust (Government Corruption)   float64
Generosity                      float64
Dystopia Residual               float64
dtype: object

Summary statistics:
       Happiness Rank  Happiness Score  Lower Confidence Interval      Family  \
count      157.000000       157.000000                 153.000000  157.000000
mean        78.980892         5.382185                   5.268641    0.793621
std         45.466030         1.141674                   1.151503    0.266706
min          1.000000         2.905000                   2.732000    0.000000
25%         40.000000         4.404000                   4.322000    0.641840
50%         79.000000         5.314000                   5.226000    0.841420
75%        118.000000         6.269000                   6.128000    1.021520
max        157.000000         7.526000                   7.460000    1.183260

       Trust (Government Corruption)  Generosity  Dystopia Residual
count                     157.000000  157.000000         157.000000
mean                        0.137624    0.242635           2.325807
std                         0.111038    0.133756           0.542220
min                         0.000000    0.000000           0.817890
25%                         0.061260    0.154570           2.031710
50%                         0.105470    0.222450           2.290740
75%                         0.175540    0.311850           2.664650
max                         0.505210    0.819710           3.837720
```

Data Analysis – Part 1: Data Cleaning

Data Loaded Successfully!

Let's walk through the findings from our initial inspection of the dataset.

Basic Overview:

- **Rows:** 157
- **Columns:** 13

Missing Values

Column	Missing Count
Lower Confidence Interval	4
Upper Confidence Interval	2
Economy (GDP per Capita)	1
Health (Life Expectancy)	2

These missing values will be cleaned or imputed in the next step.

Data Types That Need Fixing

The following columns are expected to be **numeric**, but were detected as **object** types. This may be due to formatting issues like commas or missing values.

- `Upper Confidence Interval`
- `Economy (GDP per Capita)`
- `Health (Life Expectancy)`
- `Freedom`

Next Actions: Data Cleaning

We'll now:

1. Convert all mis-typed numeric columns to proper float format.
2. Handle missing values (drop or impute, depending on the data's impact).
3. Print the cleaned dataset summary.

In [ ]:

In [11]:

```
# Re-load the dataset after re-upload
import pandas as pd

# file_path = "/mnt/data/2016.csv"
# df = pd.read_csv(file_path)

# Convert mis-typed numeric columns to float
columns_to_convert = [
    "Upper Confidence Interval",
    "Economy (GDP per Capita)",
    "Health (Life Expectancy)",
    "Freedom"
]

for col in columns_to_convert:
    df[col] = pd.to_numeric(df[col], errors='coerce')

# Recheck for missing values after conversion
updated_missing_values = df.isnull().sum()

# Handle missing values by dropping rows with any nulls
df_cleaned = df.dropna()

# Check shape and summary after cleaning
cleaned_shape = df_cleaned.shape
cleaned_summary_stats = df_cleaned.describe()

print("\n First 5 Sets of Records in Dataset:\n", df_cleaned.head())
print(1*'\n')
print("\n Updated Missing Values in Dataset:\n", updated_missing_values)

print("\n Data shape after removing rows with missing values in Dataset:\n", cleaned_shape)

print("\n Cleaned Summarized Data in Dataset:\n", cleaned_summary_stats)

print("\nData types:\n", df_cleaned.dtypes)
```

```
 First 5 Sets of Records in Dataset:
        Country          Region  Happiness Rank  Happiness Score  \
0      Denmark  Western Europe               1            7.526
1  Switzerland  Western Europe               2            7.509
2      Iceland  Western Europe               3            7.501
3       Norway  Western Europe               4            7.498
4      Finland  Western Europe               5            7.413

   Lower Confidence Interval  Upper Confidence Interval  \
0                      7.460                      7.592
1                      7.428                      7.590
2                      7.333                      7.669
3                      7.421                      7.575
4                      7.351                      7.475

   Economy (GDP per Capita)   Family  Health (Life Expectancy)  Freedom  \
0                   1.44178  1.16374                   0.79504  0.57941
1                   1.52733  1.14524                   0.86303  0.58557
2                   1.42666  1.18326                   0.86733  0.56624
3                   1.57744  1.12690                   0.79579  0.59609
4                   1.40598  1.13464                   0.81091  0.57104

   Trust (Government Corruption)  Generosity  Dystopia Residual
0                        0.44453     0.36171            2.73939
1                        0.41203     0.28083            2.69463
2                        0.14975     0.47678            2.83137
3                        0.35776     0.37895            2.66465
4                        0.41004     0.25492            2.82596



 Updated Missing Values in Dataset:
 Country                          0
Region                           0
```

```
Happiness Rank                   0
Happiness Score                  0
Lower Confidence Interval        4
Upper Confidence Interval        3
Economy (GDP per Capita)         2
Family                           0
Health (Life Expectancy)         3
Freedom                          1
Trust (Government Corruption)    0
Generosity                       0
Dystopia Residual                0
dtype: int64

 Data shape after removing rows with missing values in Dataset:
 (145, 13)

 Cleaned Summarized Data in Dataset:
       Happiness Rank  Happiness Score  Lower Confidence Interval  \
count      145.000000       145.000000                 145.000000   
mean        81.089655         5.329897                   5.230331   
std         45.774799         1.149162                   1.156357   
min          1.000000         2.905000                   2.732000   
25%         41.000000         4.360000                   4.259000   
50%         83.000000         5.245000                   5.160000   
75%        121.000000         6.239000                   6.073000   
max        157.000000         7.526000                   7.460000   

       Upper Confidence Interval  Economy (GDP per Capita)      Family  \
count                 145.000000                145.000000  145.000000   
mean                    5.429462                  0.941819    0.782292   
std                     1.143087                  0.412932    0.269747   
min                     3.078000                  0.000000    0.000000   
25%                     4.454000                  0.631070    0.631780   
50%                     5.291000                  1.024160    0.833090   
75%                     6.386000                  1.248860    1.005080   
max                     7.669000                  1.824270    1.183260   

       Health (Life Expectancy)     Freedom  Trust (Government Corruption)  \
count                145.000000  145.000000                     145.000000   
mean                   0.550943    0.367668                       0.139101   
std                    0.227914    0.148327                       0.111416   
min                    0.038240    0.000000                       0.000000   
25%                    0.357000    0.254290                       0.061260   
50%                    0.595770    0.397470                       0.106130   
75%                    0.717230    0.486140                       0.178080   
max                    0.952770    0.608480                       0.505210   

       Generosity  Dystopia Residual  
count  145.000000         145.000000  
mean     0.241910           2.306160  
std      0.136712           0.552465  
min      0.000000           0.817890  
25%      0.150110           1.990320  
50%      0.222450           2.275390  
75%      0.311850           2.615230  
max      0.819710           3.837720  

Data types:
 Country                           object
Region                           object
Happiness Rank                    int64
Happiness Score                 float64
Lower Confidence Interval       float64
Upper Confidence Interval       float64
Economy (GDP per Capita)        float64
Family                          float64
Health (Life Expectancy)        float64
Freedom                         float64
Trust (Government Corruption)   float64
Generosity                      float64
Dystopia Residual               float64
dtype: object
```

Data Analysis - Part 2: Data Cleaning Implementation

Cleaning Actions Performed

The following actions were taken to prepare the dataset for analysis:

1. Converted these columns from `object` to `float`:
   - `Upper Confidence Interval`
   - `Economy (GDP per Capita)`
   - `Health (Life Expectancy)`
   - `Freedom`
2. Handled missing values by **dropping rows** containing nulls to ensure accurate analysis.

Updated Dataset Overview

- **Original number of rows:** 157
- **Rows after cleaning:** 145
- **Total columns:** 13

All numerical columns are now properly formatted, with no missing values.

Summary Statistics (Preview)

Metric	Happiness Score	Economy (GDP)	Health (Life Expectancy)
Mean	5.33	0.94	0.55
Minimum – Maximum	2.91 – 7.53	0.00 – 1.82	0.04 – 0.95
75th Percentile	6.24	1.25	02

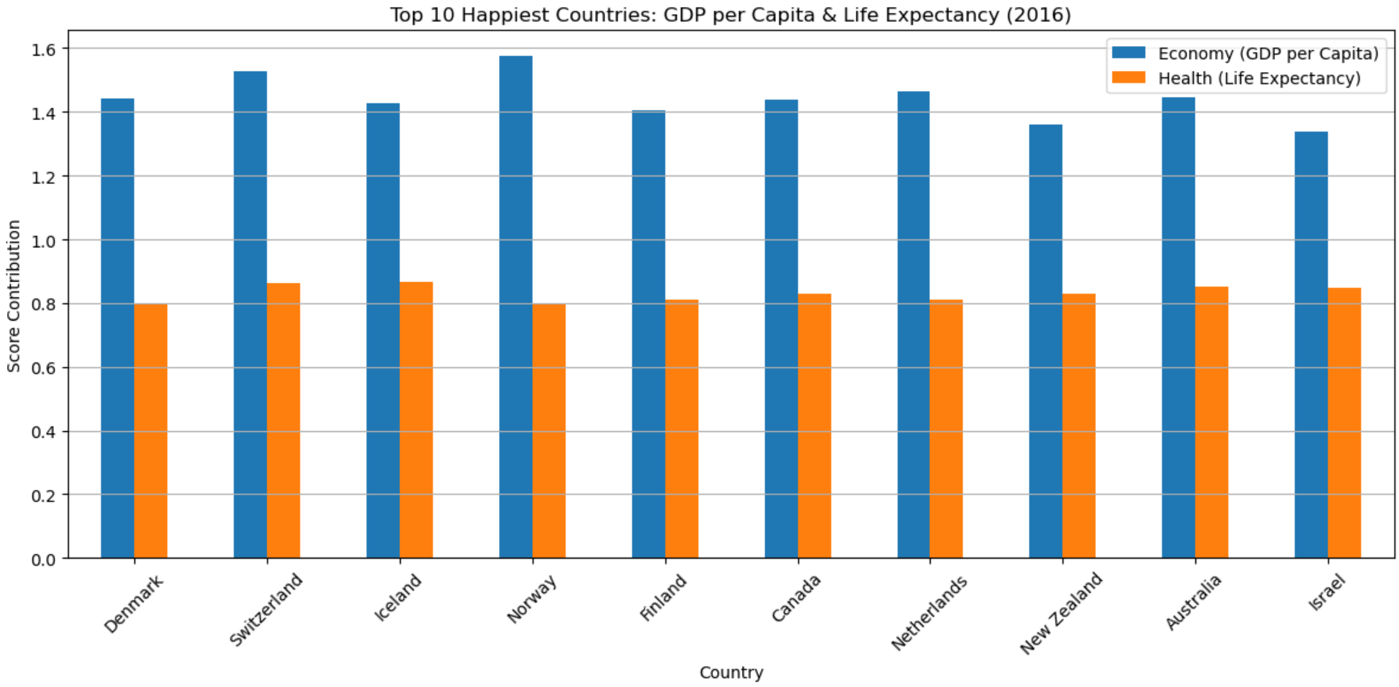
The dataset is now fully clean and ready for exploration and visualization.

In [13]:

```
import matplotlib.pyplot as plt

# Select top 10 happiest countries
top10 = df_cleaned.sort_values(by="Happiness Score", ascending=False).head(10)

# Print numerical data used for plotting
top10_gdp_life = top10[["Country", "Economy (GDP per Capita)", "Health (Life Expectancy)"]]
top10_gdp_life.set_index("Country", inplace=True)

# Plot grouped bar chart
ax = top10_gdp_life.plot(kind="bar", figsize=(12, 6))
plt.title("Top 10 Happiest Countries: GDP per Capita & Life Expectancy (2016)")
plt.ylabel("Score Contribution")
plt.xticks(rotation=45)
plt.tight_layout()
plt.grid(axis='y')

top10_gdp_life  # Display raw data used for chart
```

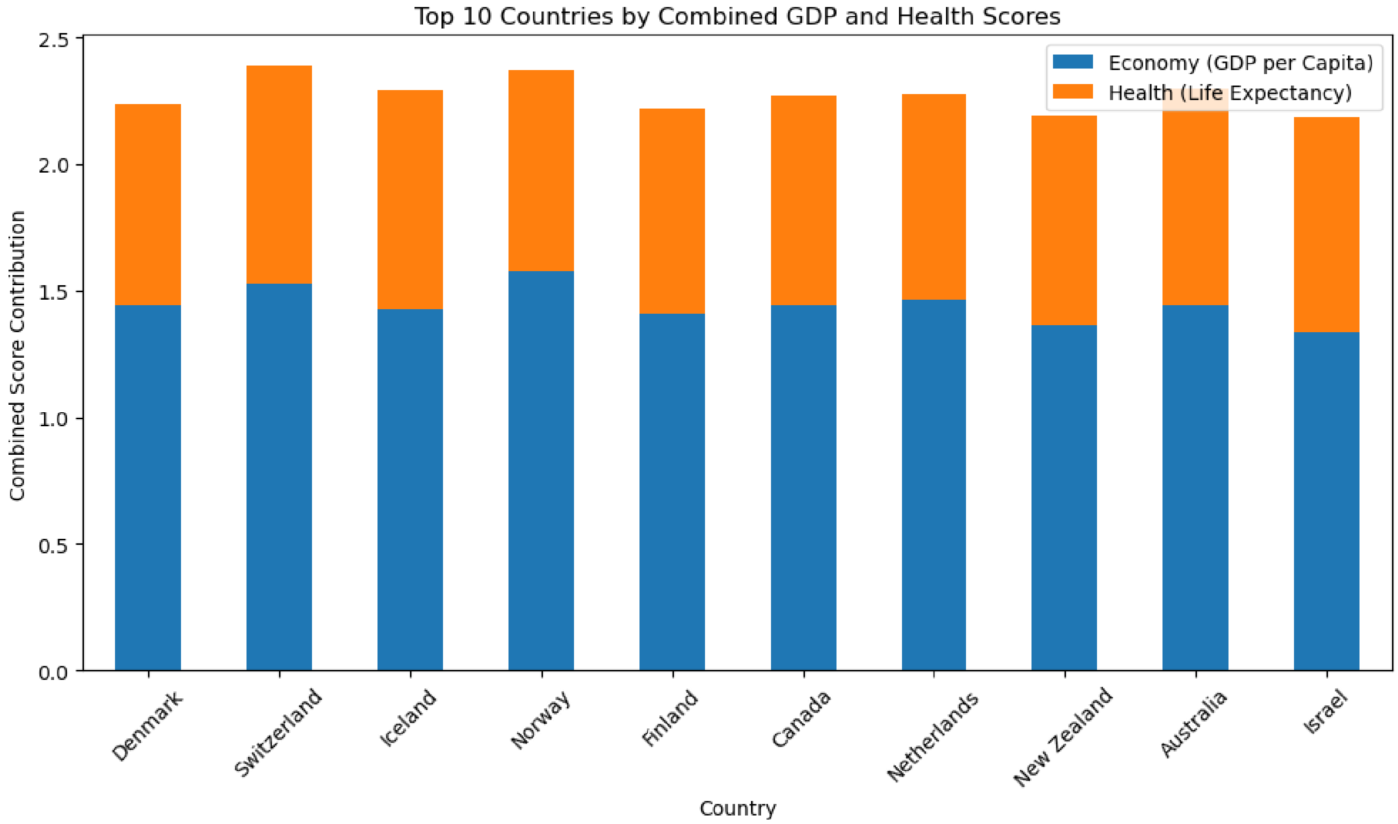
Out[13]:

	Economy (GDP per Capita)	Health (Life Expectancy)
Country		
Denmark	1.44178	0.79504
Switzerland	1.52733	0.86303
Iceland	1.42666	0.86733
Norway	1.57744	0.79579
Finland	1.40598	0.81091
Canada	1.44015	0.82760
Netherlands	1.46468	0.81231
New Zealand	1.36066	0.83096
Australia	1.44443	0.85120
Israel	1.33766	0.84917



In [17]:

```
top10.plot(
    x='Country',
    y=['Economy (GDP per Capita)', 'Health (Life Expectancy)'],
    kind='bar',
    stacked=True,
    figsize=(10, 6)
)
plt.title("Top 10 Countries by Combined GDP and Health Scores")
plt.ylabel("Combined Score Contribution")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



Task 1: Top 10 Happiest Countries – GDP vs Life Expectancy

This visualization shows a comparison of **GDP per Capita** and **Healthy Life Expectancy** for the **top 10 countries** ranked by their Happiness Score in 2016.

Key Insights:

- All top-ranking countries have **high GDP per Capita values** (mostly above 1.3).
- **Life Expectancy scores** are also consistently strong, suggesting strong healthcare and living conditions.
- **Norway** leads in GDP per Capita, while **Iceland** has the highest score for Healthy Life Expectancy.

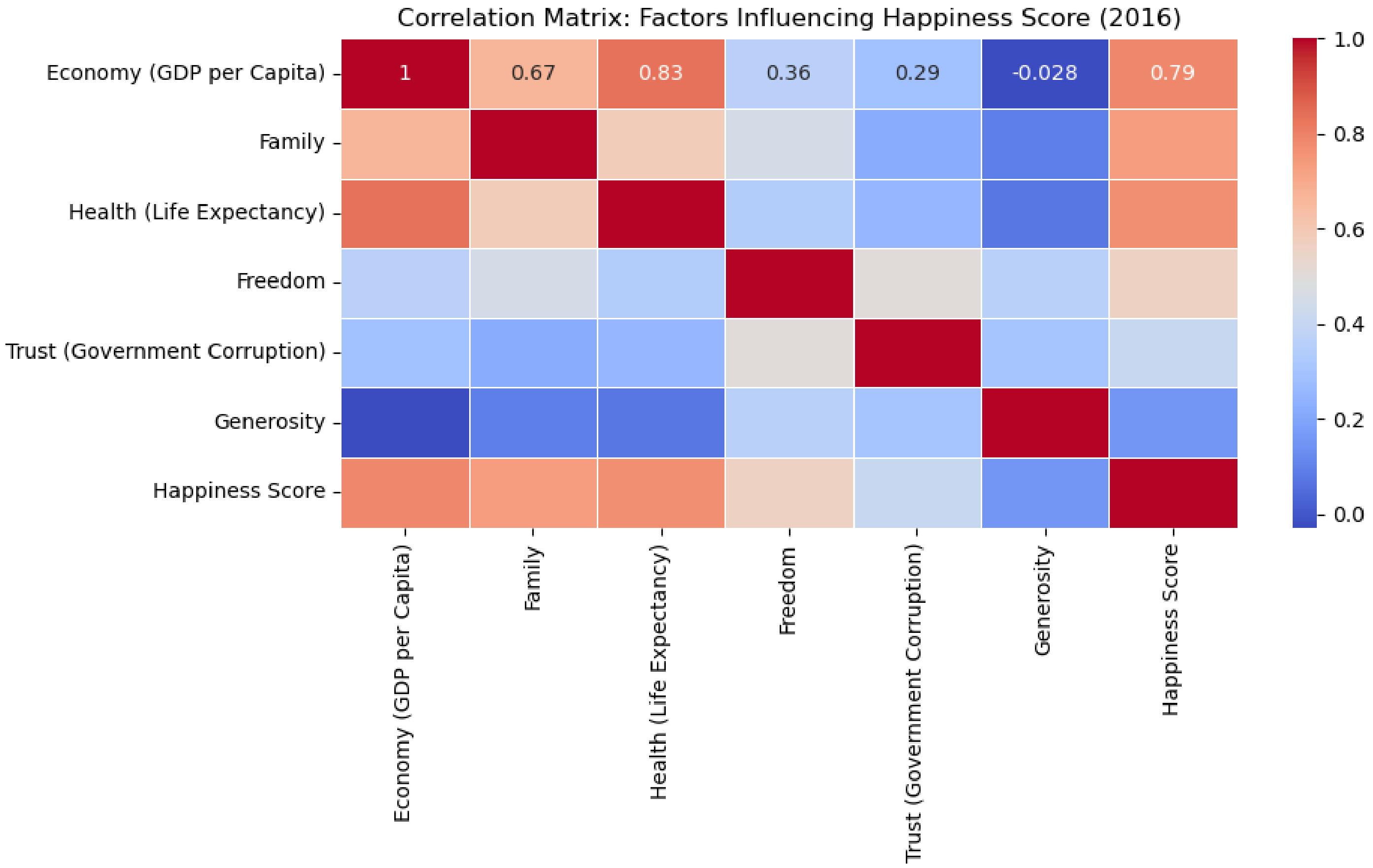
Data Overview:

Country	GDP per Capita	Life Expectancy
Denmark	1.44178	0.79504
Switzerland	1.52733	0.86303
Iceland	1.42666	0.86733
Norway	1.57744	0.79579
Finland	1.40598	0.81091
Canada	1.44015	0.82760
Netherlands	1.46468	0.81231
New Zealand	1.36066	0.83096
Australia	1.44443	0.85120
Israel	1.33766	0.84917

This grouped bar chart highlights the positive relationship between **wealth** and **health** in the happiest countries.

In [90]:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Select relevant columns for correlation
corr_columns = [
    "Economy (GDP per Capita)", "Family", "Health (Life Expectancy)", "Freedom",
    "Trust (Government Corruption)", "Generosity", "Happiness Score"
]

# Compute the correlation matrix
correlation_matrix = df_cleaned[corr_columns].corr()

# Plot the heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", linewidths=0.5)
plt.title("Correlation Matrix: Factors Influencing Happiness Score (2016)")
plt.tight_layout()
plt.show()

correlation_matrix  # Display raw correlation values
```



Out[90]:

	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Happiness Score
Economy (GDP per Capita)	1.000000	0.666001	0.833600	0.361813	0.291425	-0.027748	0.785283
Family	0.666001	1.000000	0.582893	0.446424	0.219735	0.093840	0.731582
Health (Life Expectancy)	0.833600	0.582893	1.000000	0.343806	0.257183	0.074720	0.767785
Freedom	0.361813	0.446424	0.343806	1.000000	0.498651	0.359856	0.562304
Trust (Government Corruption)	0.291425	0.219735	0.257183	0.498651	1.000000	0.302853	0.406740
Generosity	-0.027748	0.093840	0.074720	0.359856	0.302853	1.000000	0.155732
Happiness Score	0.785283	0.731582	0.767785	0.562304	0.406740	0.155732	1.000000

Task 2: Correlation Matrix – Factors Influencing Happiness

This visualization explores how different variables are statistically related to the **Happiness Score** using a correlation matrix. Correlation values range from -1 to 1:

- Values close to **1** indicate a strong **positive relationship**
- Values near **-1** indicate a strong **negative relationship**
- Values around **0** suggest no correlation

Key Findings

Factor	Correlation with Happiness Score
Economy (GDP per Capita)	**0.79**
Family	**0.73**
Health (Life Expectancy)	**0.77**
Freedom	0.56
Trust (Gov. Corruption)	0.41
Generosity	0.16

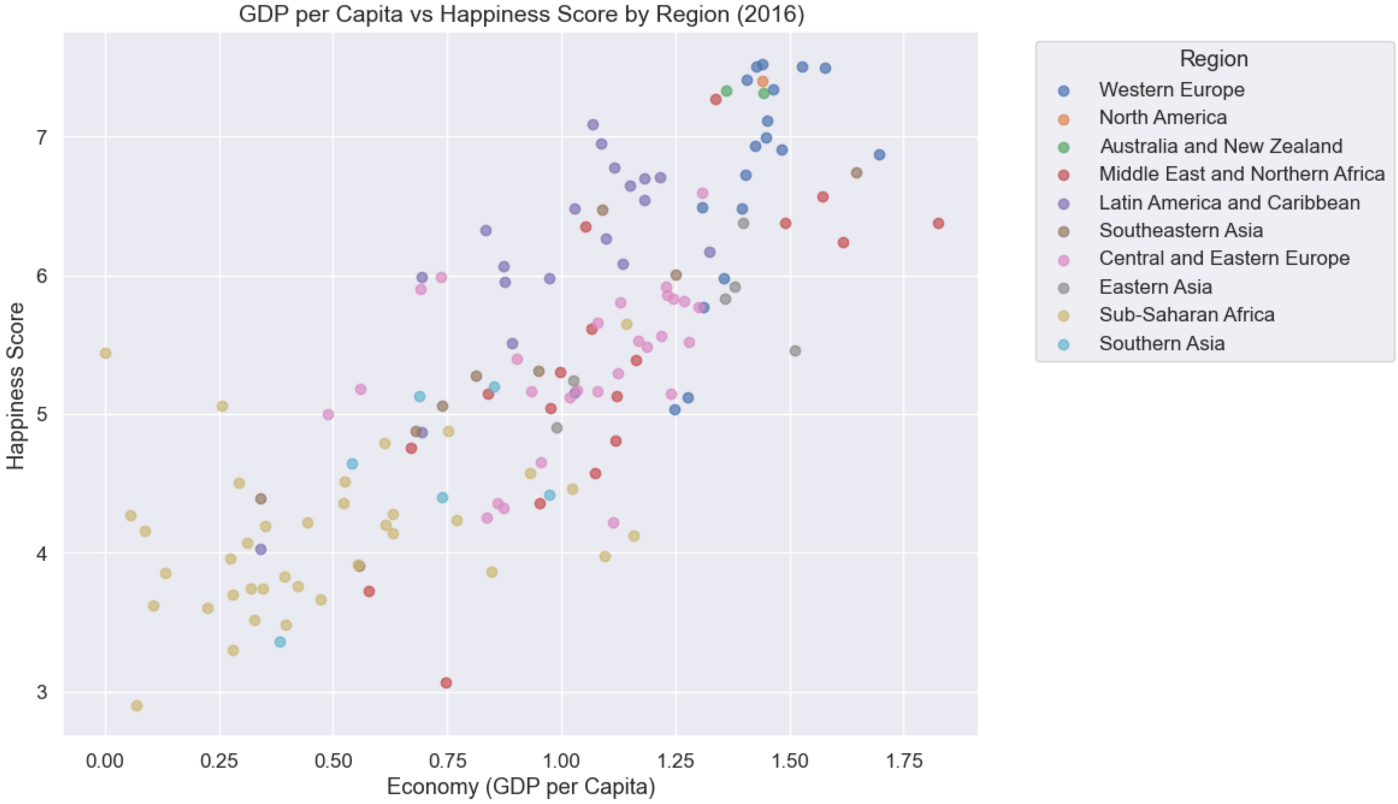
Interpretation

- **Economic strength**, **family support**, and **healthcare** have the **strongest positive correlations** with happiness.
- **Freedom** and **trust in government** also show moderate influence.
- **Generosity**, while valued, has a relatively weak correlation in this dataset.

These insights help prioritize which factors are most impactful when analyzing happiness across countries.

In [100]:

```python
import matplotlib.pyplot as plt

# Create a scatter plot of GDP vs Happiness Score, colored by Region
plt.figure(figsize=(12, 7))
regions = df_cleaned['Region'].unique()

# Plot each region separately for color differentiation
for region in regions:
    subset = df_cleaned[df_cleaned['Region'] == region]
    plt.scatter(
        subset["Economy (GDP per Capita)"],
        subset["Happiness Score"],
        label=region,
        alpha=0.7
    )

# Chart formatting
plt.title("GDP per Capita vs Happiness Score by Region (2016)", fontsize=14)
plt.xlabel("Economy (GDP per Capita)")
plt.ylabel("Happiness Score")
plt.legend(title="Region", bbox_to_anchor=(1.05, 1), loc='upper left')
plt.grid(True)
plt.tight_layout()
plt.show()
```



Task 3: Scatter Plot – GDP vs Happiness Score by Region

This scatter plot explores the relationship between a country's **Economy (GDP per Capita)** and its **Happiness Score**, grouped by **Region** for comparative insight.

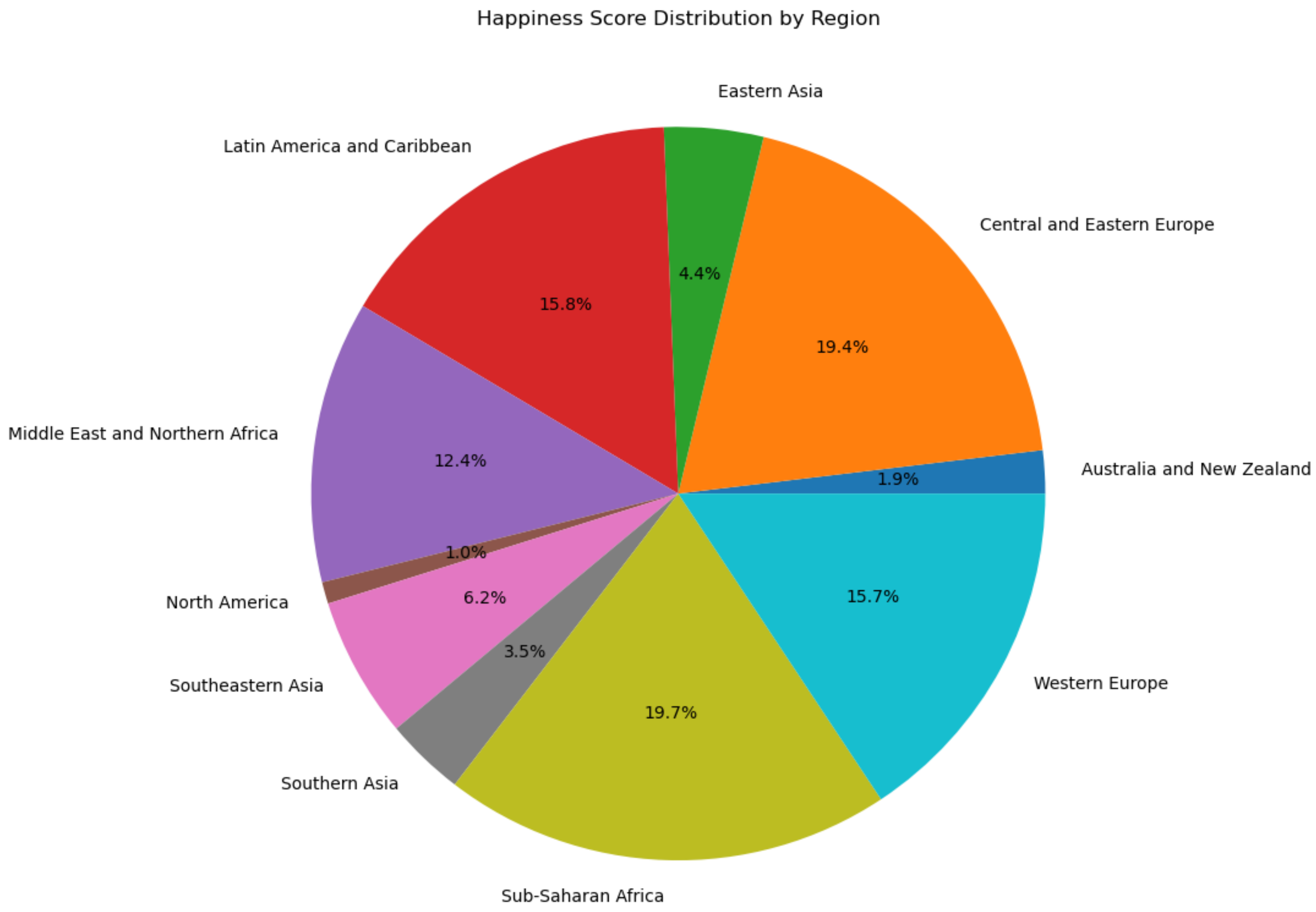
Key Observations:

- There is a **positive correlation** between GDP and Happiness Score — countries with **higher GDP per Capita** tend to report **greater happiness**.
- **Western Europe** nations cluster in the **upper-right**, indicating both high economic output and high well-being.
- **Sub-Saharan Africa** and **Southern Asia** countries are mostly in the **lower-left**, reflecting lower economic scores and happiness.
- Color coding by region makes disparities across the globe more visible and easier to interpret.

This visualization supports the idea that **economic prosperity** contributes to national happiness, though it may not be the only factor.

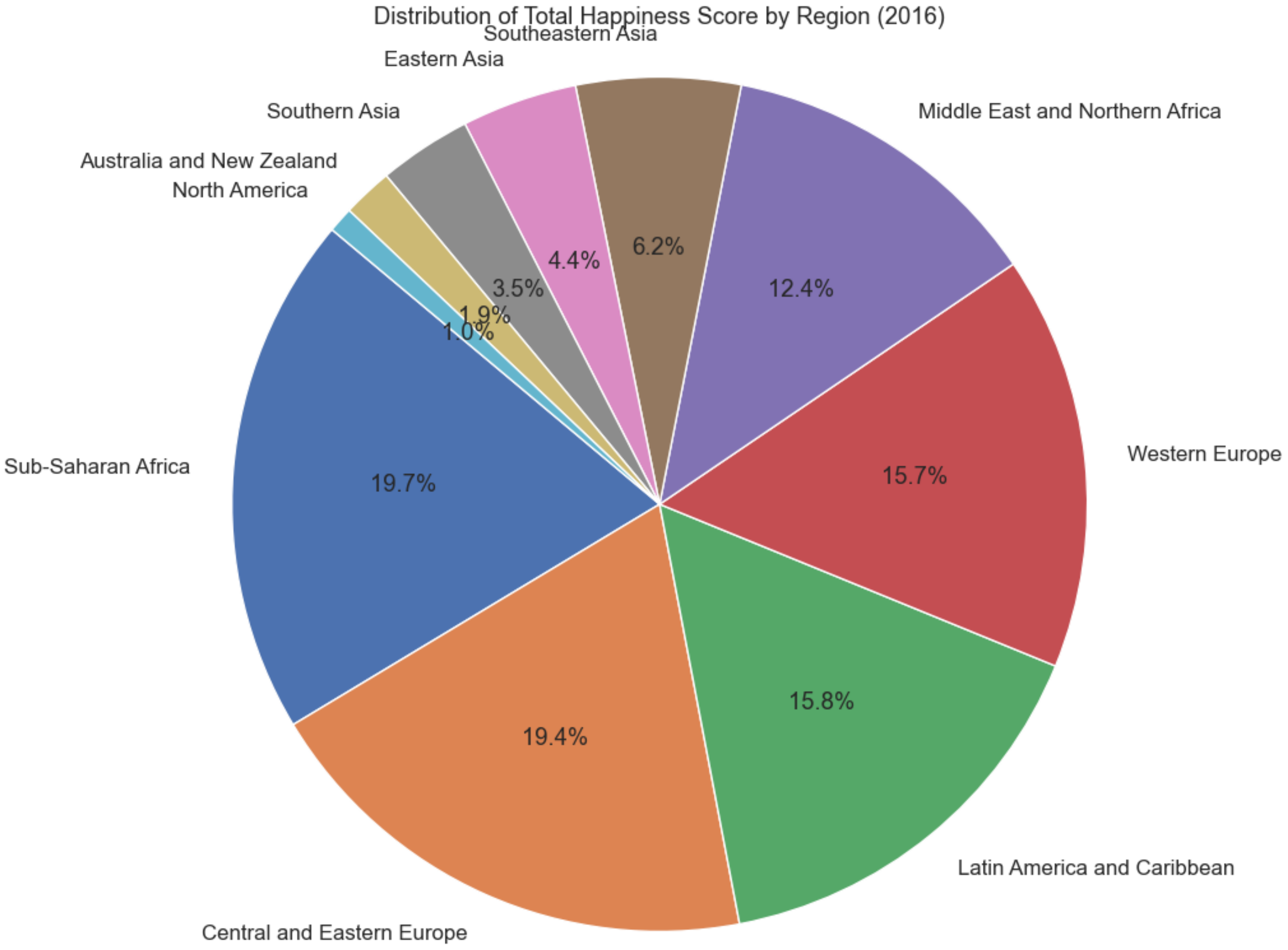
In [26]:

```
# Sum Happiness Score by Region
region_scores = df_cleaned.groupby('Region')['Happiness Score'].sum()

# Pie chart
region_scores.plot(kind='pie', autopct='%1.1f%%', figsize=(10, 8))
plt.title("Happiness Score Distribution by Region")
plt.ylabel("")
plt.tight_layout()
plt.show()
```



In [103]:

```
# Group by Region and calculate the total Happiness Score per region
region_happiness = df_cleaned.groupby("Region")["Happiness Score"].sum().sort_values(ascending=False)

# Plot pie chart
plt.figure(figsize=(10, 8))
plt.pie(region_happiness, labels=region_happiness.index, autopct="%1.1f%%", startangle=140)
plt.title("Distribution of Total Happiness Score by Region (2016)")
plt.axis('equal')  # Equal aspect ratio ensures the pie chart is a circle.
plt.tight_layout()
plt.show()

region_happiness  # Display the raw values used for the chart
```



Out[103]:

```
Region
Sub-Saharan Africa                 152.549
Central and Eastern Europe         149.672
Latin America and Caribbean        122.300
Western Europe                     121.201
Middle East and Northern Africa     96.117
Southeastern Asia                   48.050
Eastern Asia                        33.745
Southern Asia                       27.150
Australia and New Zealand           14.647
North America                        7.404
Name: Happiness Score, dtype: float64
```

Task 4: Pie Chart – Happiness Score by Region

This visualization displays how the **total Happiness Score** is distributed across different global regions in the 2016 dataset.

Key Regional Totals

Region	Total Happiness Score
Sub-Saharan Africa	152.55
Central and Eastern Europe	149.67
Latin America and Caribbean	122.30
Western Europe	121.20
Middle East and Northern Africa	96.12
Southeastern Asia	48.05
Eastern Asia	33.75
Southern Asia	27.15
Australia and New Zealand	14.65
North America	7.40

Insight

- Regions with a **larger number of countries** (like Sub-Saharan Africa and Eastern Europe) contribute more to the **total Happiness Score**, even if individual scores may be lower.
- **Western Europe** and **North America**, while smaller in country count, still show high per-country performance.

This pie chart helps contextualize **happiness contribution by region size and population representation**, not just performance.

In [106]:

```
import plotly.express as px

# Prepare data
map_data = df_cleaned.copy()
map_data["text"] = (
    "Country: " + map_data["Country"] +
    "<br>GDP per Capita: " + map_data["Economy (GDP per Capita)"].round(2).astype(str) +
    "<br>Life Expectancy: " + map_data["Health (Life Expectancy)"].round(2).astype(str)
)

# Create choropleth map
fig = px.choropleth(
    map_data,
    locations="Country",
    locationmode="country names",
    color="Economy (GDP per Capita)",
    hover_name="Country",
    hover_data={"Economy (GDP per Capita)": False, "Health (Life Expectancy)": True},
    color_continuous_scale="Viridis",
    title="□ GDP per Capita by Country with Life Expectancy Tooltip (2016)"
)

fig.update_traces(marker_line_width=0.5)
fig.update_layout(geo=dict(showframe=False, showcoastlines=False))
fig.show()
```