World Happiness Report - Analysis

Introduction

For my data analysis, I decided to examine data from the **World Happiness Report from 2015 to 2023**. It is based on a global survey in which the population rates their own lives on a scale from 0 to 10. The reason I chose this data is that there are several key factors that influence this value. In my analysis, these values are as follows:

1. GDP per capita
2. Social support
3. Healthy life expectancy
4. Freedom to make life choices
5. Generosity
6. Perceptions of corruption

The data comes from various sources/years and is divided into different CSV files. These are getting merged and analyzed.

Research Questions

1. Which variables have the strongest influence on the happiness score?
2. Has the coronavirus pandemic had an impact on the happiness score?
3. To what extent can a country's happiness score be predicted using a Linear Regression model?
4. Are there regional differences in terms of the happiness score?

Key limitations

There are limitations in my data research that affect the quality of the data.

On the one hand, the perception of one's own well-being is very subjective. This means that a certain number of points can be interpreted differently (either as good or bad). If one person rates their quality of life as a "7," the same quality of life may be a "5" for another person. In addition, this data consists of six factors. The problem behind this is that important data is missing, such as environmental quality or inequality. **I will address other limitations in the course of my research.**

Data Preparation and Merging

To meet the requirements of my data analysis, I will first merge all tables. I do this by adding a new column in which I insert the year of the data record. The result is a CSV file with 1367 rows.

In [16]:
```
import pandas as pd
import glob
import os
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

file_paths = sorted(glob.glob("WHR_*.csv"))
all_dfs = []

for file in file_paths:
    temp_df = pd.read_csv(file)

    year_val = int(os.path.basename(file).split('_')[1].split('.')[0])
    temp_df['Year'] = year_val
    all_dfs.append(temp_df)


df = pd.concat(all_dfs, ignore_index=True)
df = df.fillna(df.mean(numeric_only=True))

display(df)
```

	country	region	happiness_score	gdp_per_capita	social_support	healthy_life_expectancy	freedom_to_make_life_choices	generosity	perceptions_of_corruption	Year
0	Switzerland	Western Europe	7.587	1.39651	1.34951	0.94143	0.66557	0.29678	0.41978	2015
1	Iceland	Western Europe	7.561	1.30232	1.40223	0.94784	0.62877	0.43630	0.14145	2015
2	Denmark	Western Europe	7.527	1.32548	1.36058	0.87464	0.64938	0.34139	0.48357	2015
3	Norway	Western Europe	7.522	1.45900	1.33095	0.88521	0.66973	0.34699	0.36503	2015
4	Canada	North America and ANZ	7.427	1.32629	1.32261	0.90563	0.63297	0.45811	0.32957	2015
...	...	...	...	...	...	...	...	...	...	...
1362	Congo (Kinshasa)	Sub-Saharan Africa	3.207	0.53100	0.78400	0.10500	0.37500	0.18300	0.06800	2023
1363	Zimbabwe	Sub-Saharan Africa	3.204	0.75800	0.88100	0.06900	0.36300	0.11200	0.11700	2023
1364	Sierra Leone	Sub-Saharan Africa	3.138	0.67000	0.54000	0.09200	0.37100	0.19300	0.05100	2023
1365	Lebanon	Middle East and North Africa	2.392	1.41700	0.47600	0.39800	0.12300	0.06100	0.02700	2023
1366	Afghanistan	South Asia	1.859	0.64500	0.00000	0.08700	0.00000	0.09300	0.05900	2023

1367 rows × 10 columns

In [17]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1367 entries, 0 to 1366
Data columns (total 10 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   country                       1367 non-null   object
 1   region                        1367 non-null   object
 2   happiness_score               1367 non-null   float64
 3   gdp_per_capita                1367 non-null   float64
 4   social_support                1367 non-null   float64
 5   healthy_life_expectancy       1367 non-null   float64
 6   freedom_to_make_life_choices  1367 non-null   float64
 7   generosity                    1367 non-null   float64
 8   perceptions_of_corruption     1367 non-null   float64
 9   Year                          1367 non-null   int64
dtypes: float64(7), int64(1), object(2)
memory usage: 106.9+ KB
```

To ensure data quality, I use **"df.info()"**. This allows me to check that no errors have occurred when merging all the data. Since I now have 10 columns (0-9), this means that the new column for specifying the individual data record (Year) is included. This is also indicated in the output under "9". The so-called "non-null count" is also very relevant. This ensures that my merged file does not contain any missing data. To guarantee this, I first used "df = df.fillna(df.mean(numeric_only=True))" to fill in all missing values with the calculated average value. This allowed me to significantly improve my data quality, as otherwise my regression models would not work (NaN throws an error code). In addition, I can identify the data types of the individual columns (under "Dtype"). "Float64" stands for decimal numbers (happiness_score, gdp_per_capita...), "object" usually stands for text (country, region), and "int64" stands for integers (Year). Last but not least, you can see how much memory is used by the data. In my example, it's only a little over 100KB, but with large amounts of data, hardware performance can suffer.

General Statistic (Distribution and Average Happiness Score over the Years) + Key questions

The main topic of my data analysis is the "Happiness Score." Accordingly, I decided to take a closer look at this value first. The lower plots show the frequencies of all values of the happiness_score. From this, I can draw the following conclusions:

Mean (AVG): The global average across all years is **5.44**
Range (min/max): The distribution shows enormous differences. While most countries lie between approximately 4.3 and 6.5, there are also those that fall significantly outside the range. The gap between happy and unhappy ranges from Afghanistan in 2023 with a value of **1.86 to 7.84**, which was recorded by Finland in 2021.

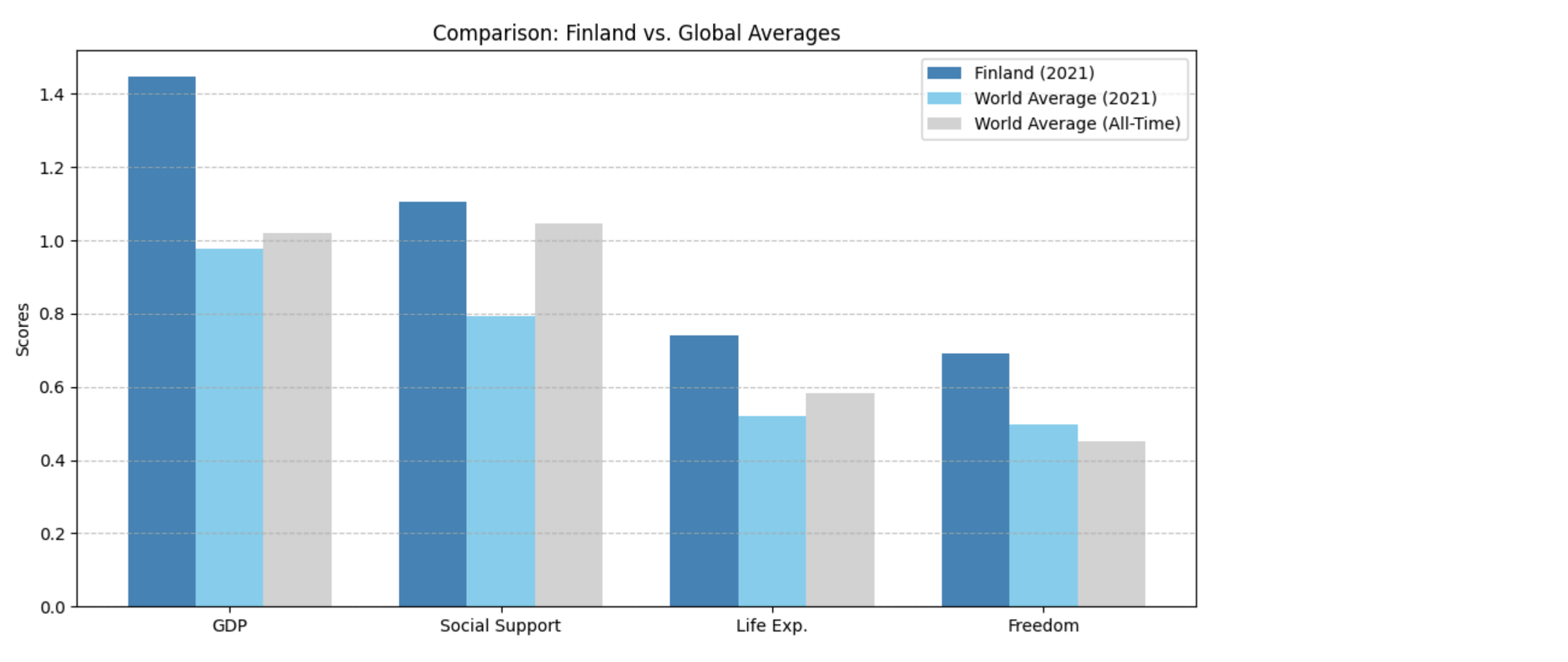
In [18]:
```
plt.figure()
sns.histplot(df['happiness_score'], kde=True)
plt.title('Statistical Distribution of Happiness Scores')
plt.xlabel('Score')
plt.ylabel('Frequency')
plt.show()

all_time_avg = df['happiness_score'].mean()
print(f"The all-time average happiness score (2015-2023) is: {all_time_avg:.2f}")

top_row = df.loc[df['happiness_score'].idxmax()]
bottom_row = df.loc[df['happiness_score'].idxmin()]

print(f"Happiest Country: {top_row['country']} ({top_row['Year']}) with a score of {top_row['happiness_score']:.2f}")
print(f"Unhappiest Country: {bottom_row['country']} ({bottom_row['Year']}) with a score of {bottom_row['happiness_score']:.2f}")
```


```
The all-time average happiness score (2015-2023) is: 5.44
Happiest Country: Finland (2021) with a score of 7.84
Unhappiest Country: Afghanistan (2023) with a score of 1.86
```

This finding became an interesting observation because such a high value was recorded during the COVID-19 pandemic. After extensive research, I came to the conclusion that a combination of social support, life expectancy, and freedom contribute to Finland's high happiness score.

In [19]:
```
display(df[(df['country'] == 'Finland') & (df['Year'] == 2021)])

features_to_plot = ['gdp_per_capita', 'social_support', 'healthy_life_expectancy', 'freedom_to_make_life_choices']


finland_2021 = df[(df['country'] == 'Finland') & (df['Year'] == 2021)][features_to_plot].iloc[0]
world_avg_2021 = df[df['Year'] == 2021][features_to_plot].mean()
world_avg_all_time = df[features_to_plot].mean()

labels = ['GDP', 'Social Support', 'Life Exp.', 'Freedom']
x = np.arange(len(labels))
width = 0.25

fig, ax = plt.subplots(figsize=(12, 6))

ax.bar(x - width, finland_2021, width, label='Finland (2021)', color='steelblue')
ax.bar(x, world_avg_2021, width, label='World Average (2021)', color='skyblue')
ax.bar(x + width, world_avg_all_time, width, label='World Average (All-Time)', color='lightgrey')

ax.set_ylabel('Scores')
ax.set_title('Comparison: Finland vs. Global Averages')
ax.set_xticks(x)
ax.set_xticklabels(labels)
ax.legend()
ax.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()
```

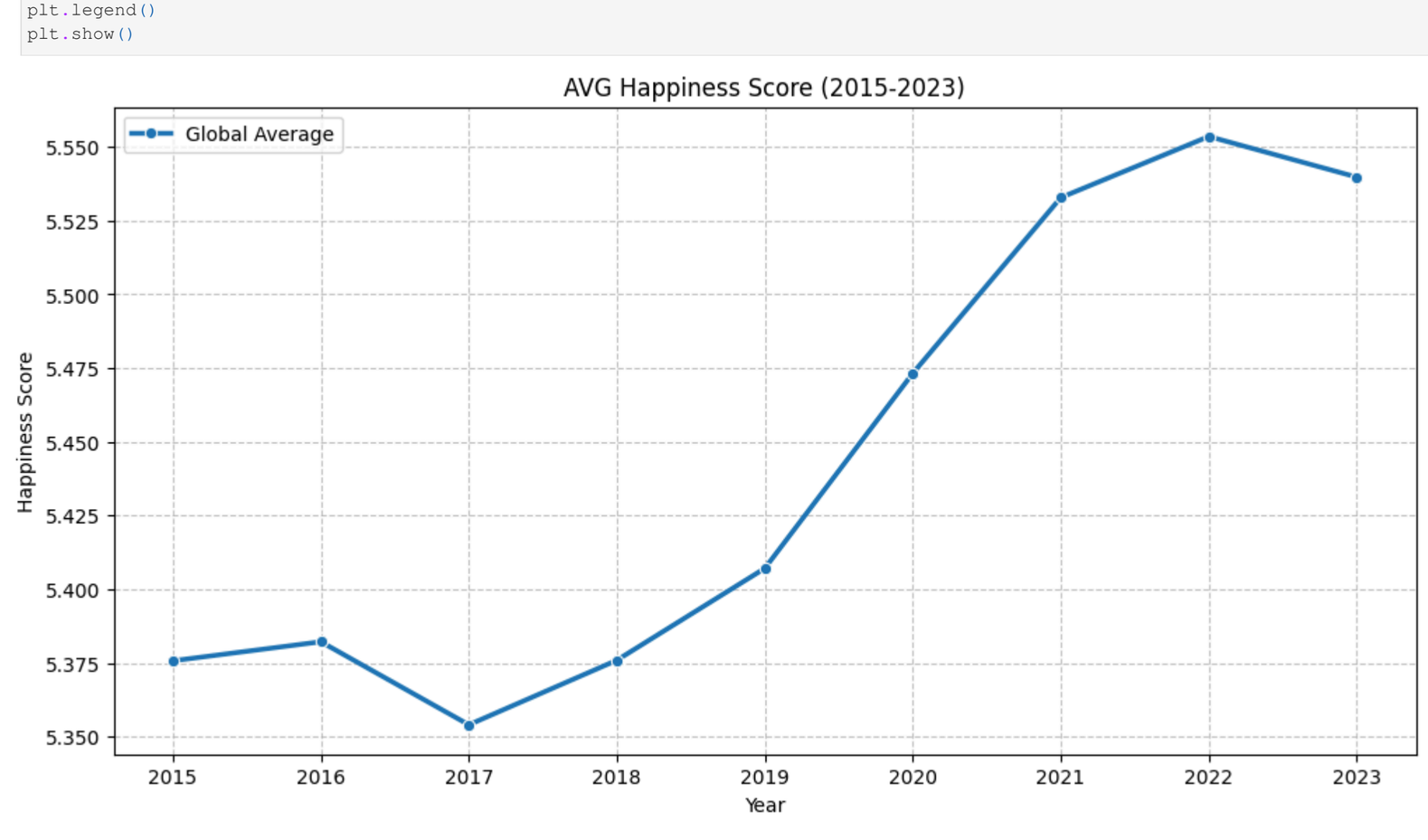
	country	region	happiness_score	gdp_per_capita	social_support	healthy_life_expectancy	freedom_to_make_life_choices	generosity	perceptions_of_corruption	Year
935	Finland	Western Europe	7.842	1.446	1.106	0.741	0.691	0.124	0.481	2021



When we look at the diagram **"comparison: Finland vs. global averages"**, we see that Finland has extremely good values. These are the key indicator for Finland's high happiness score. But this diagram raised another question for me:

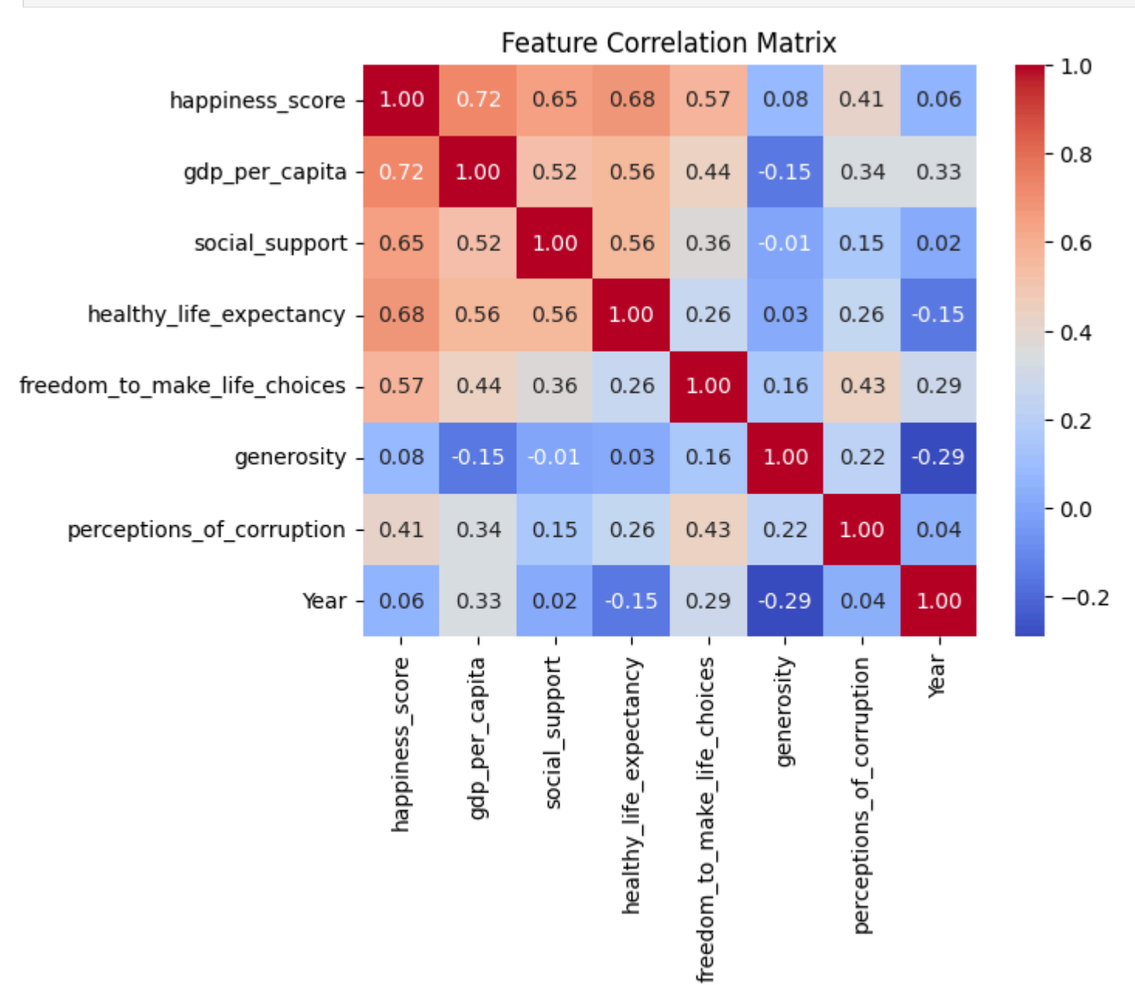
How can the "freedom_to_make_life_choices" value in 2021 (COVID-19 pandemic) be above the all-time world average?
To answer this, we need to understand what this index actually measures. This is because it is not a question of whether or not you are currently allowed to go to a restaurant, but whether the population feels that they are allowed to make important life decisions for themselves. Furthermore, trust reduces the feeling of coercion. If people consider the coronavirus measures to be sensible and necessary, they do not perceive them as a deprivation of freedom but as a protective measure.

In [20]:
```
yearly_avg = df.groupby('Year')['happiness_score'].mean().reset_index()

plt.figure(figsize=(12, 6))
sns.lineplot(data=yearly_avg, x='Year', y='happiness_score', marker='o', linewidth=2.5, label='Global Average')

plt.title('AVG Happiness Score (2015-2023)')
plt.xlabel('Year')
plt.ylabel('Happiness Score')
plt.grid(True, linestyle='--', alpha=0.7)
plt.legend()
plt.show()
```



The **"AVG Happiness Score"** graph shows the average development of the happiness score over the years. It is important to note that the year itself does not actually influence the happiness score (see causality cube below). Here, too, the development during the coronavirus pandemic is very interesting. As mentioned above, the happiness score is a subjective assessment of life, which explains this increase as follows: **People feel more grateful for their lives when they see that they and their families are surviving such a global crisis.**

In [21]:
```
plt.figure()
correlation_matrix = df.select_dtypes(include=['number']).corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Feature Correlation Matrix')
plt.show()
```



The **"Feature Correlation Matrix"** shows the so-called correlation cube. My intention is to see how the individual key attributes affect the happiness score and how the individual attributes correlate with each other. At first glance, the result seems clear:
"gdp_per_capita" appears to be the value that has the most influence on the "happiness_score." The higher the value, the higher the happiness score. Apparently, "healthy_life_expectancy," "social_support," and "freedom_to_make_life_choices" also have a positive effect on the happiness score, while the year or the generosity of the population has little effect on it. However, extreme caution is advised with these assumptions, because:

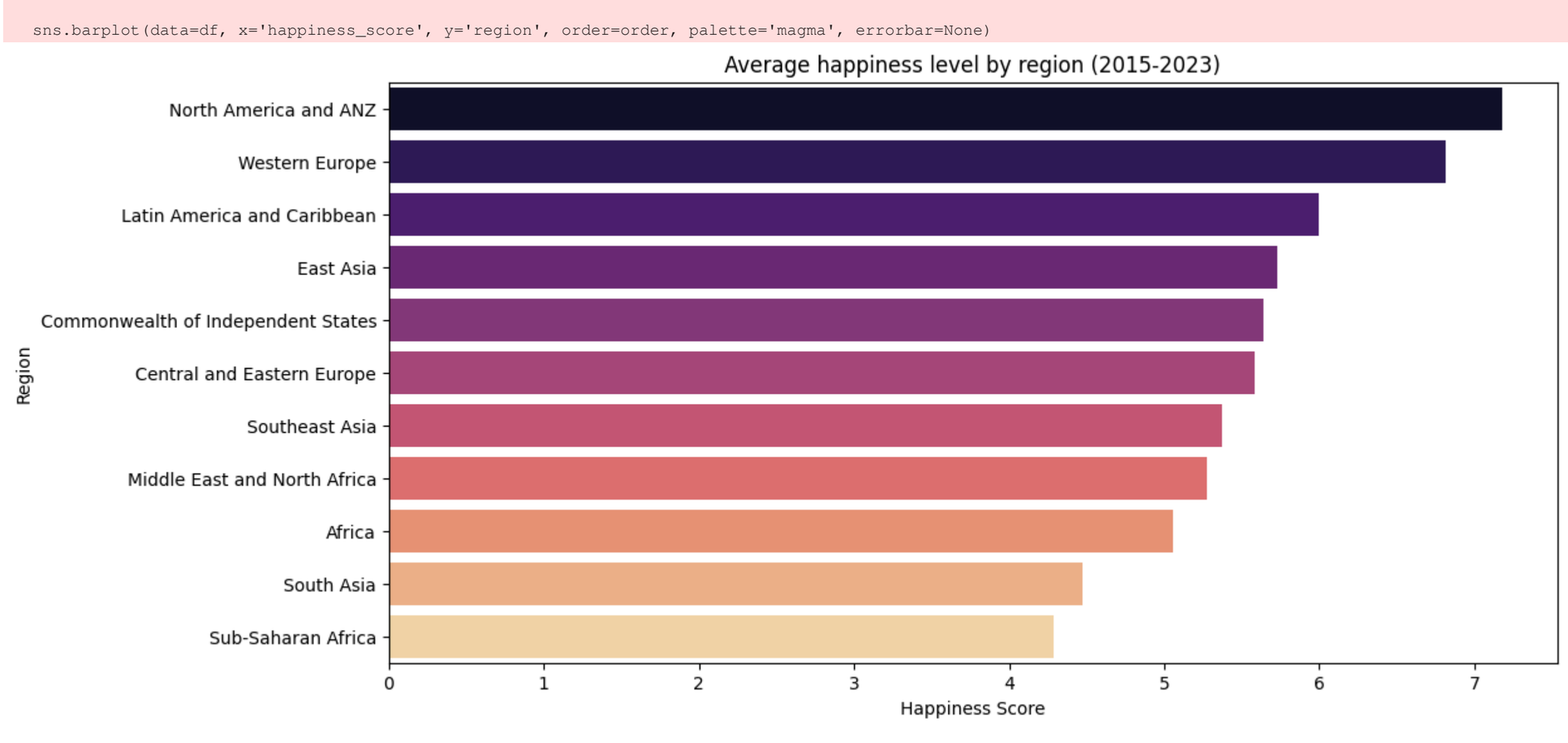
Correlation does not equal causality

How these values are related is another key limitation of my project. This becomes clear with the following example: If a country is happy and has a high GDP per capita, we assume that GDP per capita influences the happiness score. However, this may not be the case. Namely, one can also reverse the assumption by saying that the high happiness score contributes to people being happier and more motivated to go to work and strengthen the economy accordingly, which in turn has an impact on a high GDP per capita.

Another question I asked myself was whether there are regional differences in terms of happiness scores. To answer this, I decided to create a bar chart and group the average happiness scores of all countries in each "region." Again, the results should be viewed with caution, as there are countries that can significantly lower the average value of a region. In addition, some countries were grouped unfavorably. For example, while Germany, Austria, and Switzerland belong geographically to Central Europe, they were assigned to "Western Europe." Nevertheless, the overall picture is still plausible, as regions with countries that have important industries (GDP per capita), a good social system, and low corruption scores stand out positively.

In [22]:
```
plt.figure(figsize=(12, 6))
order = df.groupby('region')['happiness_score'].mean().sort_values(ascending=False).index

sns.barplot(data=df, x='happiness_score', y='region', order=order, palette='magma', errorbar=None)

plt.title('Average happiness level by region (2015-2023)')
plt.xlabel('Happiness Score')
plt.ylabel('Region')
plt.show()
```

```
C:\Users\Samu\AppData\Local\Temp\ipykernel_27280\2165014225.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(data=df, x='happiness_score', y='region', order=order, palette='magma', errorbar=None)
```



Regression Models

Explanation of the models

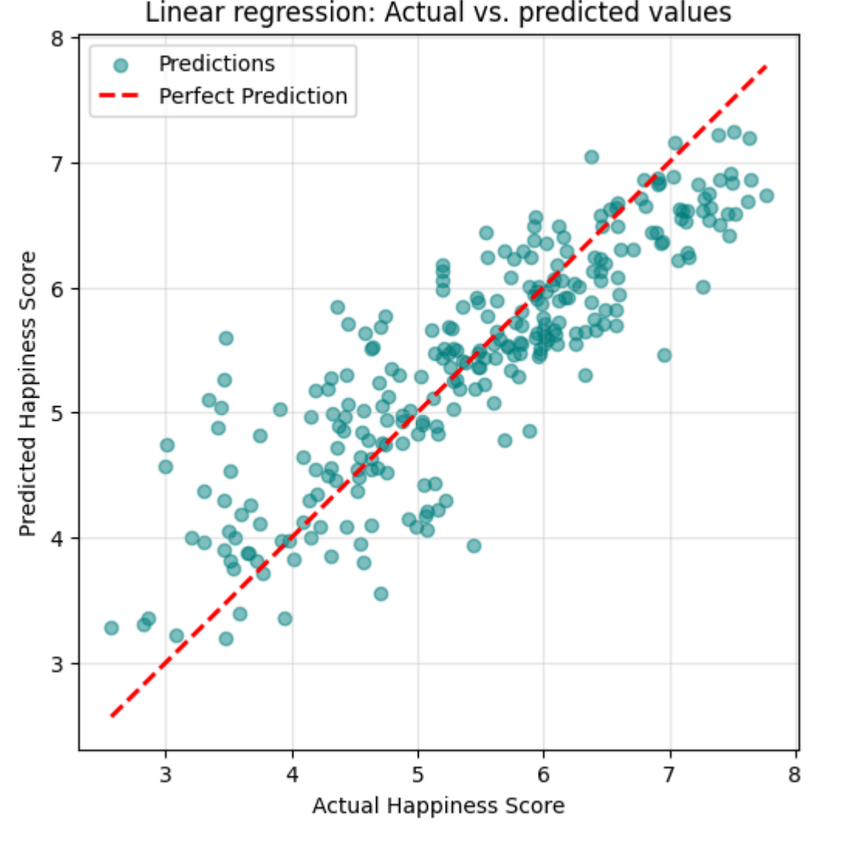
In this section, I implemented a Multiple Linear Regression model. The goal was to predict the happiness_score based on three primary features: GDP per capita, social support, and healthy life expectancy.

In [23]:
```
features = ['gdp_per_capita', 'healthy_life_expectancy', 'freedom_to_make_life_choices']
X = df[features]
y = df['happiness_score']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
print(f"R² Score: {r2_score(y_test, y_pred):.2f}")
print(f"Mean Squared Error: {mean_squared_error(y_test, y_pred):.2f}")
```

```
R² Score: 0.73
Mean Squared Error: 0.37
```

What are these values?
R^2 score: This value describes what percentage of global happiness is explained by these three values alone. In my example, I arrived at a value of 0.73, which corresponds to 73%.
Mean Squared Error: Using this value, I can evaluate how far the predictions deviate from the actual value on average. In my case, the value is 0.37. However, since the scale ranges from 0 to 10, the root must be taken from the value. This results in a standard deviation of 0.61. This means that my model is only off by about 0.6 points. This is also confirmed by my models, as a good pattern can be identified.

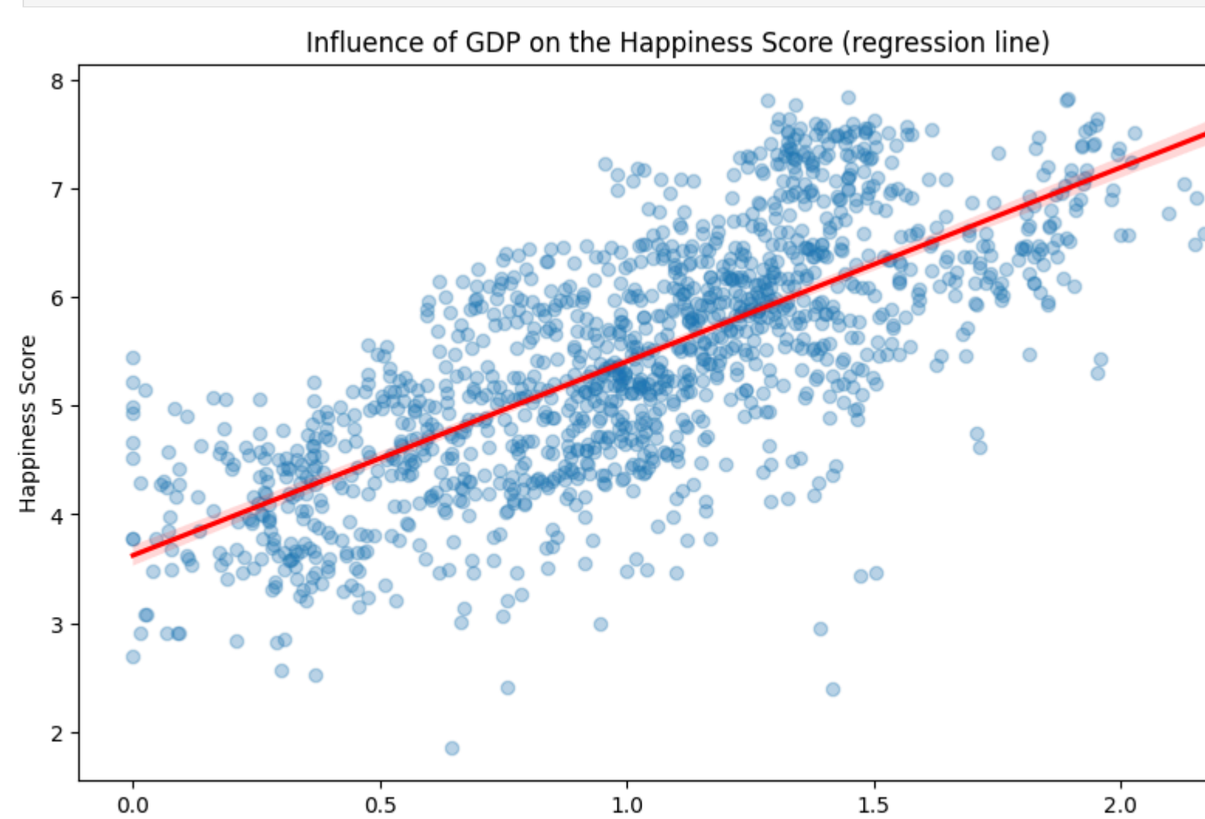
In [24]:
```
import matplotlib.pyplot as plt

plt.figure(figsize=(6, 6))
plt.scatter(y_test, y_pred, alpha=0.5, color='teal', label='Predictions')

plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()],
         color='red', lw=2, linestyle='--', label='Perfect Prediction')

plt.title('Linear regression: Actual vs. predicted values')
plt.xlabel('Actual Happiness Score')
plt.ylabel('Predicted Happiness Score')
plt.legend()
plt.grid(True, alpha=0.3)
plt.show()
```



This diagram (**"Linear regression: Actual vs. predicted values""**) is the direct result of my machine learning model. It compares the actual happiness values from my test data set with the values calculated by the model based on the features (GDP, social support, freedom).

- The red dotted line represents the perfect prediction.
- The blue dots show the performance of my linear regression.

In [25]:
```
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.regplot(data=df, x='gdp_per_capita', y='happiness_score',
            scatter_kws={'alpha':0.3}, line_kws={'color':'red'})

plt.title('Influence of GDP on the Happiness Score (regression line)')
plt.xlabel('GDP per capita')
plt.ylabel('Happiness Score')
plt.show()
```



The **"Influence of GDP on the Happiness Score (regression line)"** graph describes the relationship between GDP (the factor with the highest correlation with the happiness score) and the happiness score. Here, too, a trend can be observed:

- the red regression line shows that increasing economic performance also leads to an increase in life satisfaction.
- however, the scatter plot shows that GDP alone is not everything (see **correlation cube**). The vertical dispersion is an indicator that other factors also play a role, as otherwise the points would have to lie exactly on the regression line.

Final conclusion

I am quite satisfied with my data analysis and find the data I found to be very good for designing the project. This is because the World Happiness Index correlates with many other factors, which I believe I have illustrated well. Technically, I had the most work with the regression model. The reason for this was that my data contained empty values ("NaN"). This resulted in an error code ("ValueError: Input contains NaN") because my model cannot work with missing values. I only realized that the data integrity was minimally flawed after extensive research and troubleshooting. These were then smoothed out using "mean imputation" (already described above). I also found the development of the happiness score during the pandemic particularly exciting, as well as the fact that Finland was the happiest country in the world between 2015 and 2023, in a year marked by the coronavirus pandemic. This was not originally one of my key questions, but rather became one by chance. I found the analysis of the problem so interesting because it led me to the next hurdle: the feeling of freedom, which was above the average for all nine years during 2021.

Self-criticism

I was not entirely satisfied with the analysis of regional differences in terms of happiness scores, as this was very plausible according to my previous data analysis and did not raise any further questions for me. Accordingly, I left the diagram as it was. What's more, I could have structured my project a little better. I could also have checked the lowest happiness index to draw conclusions from it. Afghanistan has the lowest value, but since I could understand this value due to the political situation, I did not go into it further.

Data Source

The data used for this analysis is compiled from the annual World Happiness Reports (2015-2023). These reports are a publication of the Sustainable Development Solutions Network, powered by the Gallup World Poll data.

Link: https://www.kaggle.com/datasets/sazidthe1/global-happiness-scores-and-factors?resource=download&select=WHR_2016.csv