2507127_MN5813

January 13, 2025

0.1 DATA DRIVEN ANALYSIS - CORRELATION BETWEEN LIFE EXPECTANCY, SOCIAL SUPPORT AND HAPPINESS INDEX

0.1.1 PROJECT SUMMARY

This study explores the relationships between life expectancy, social support, and the happiness index across countries. The primary goal is to determine whether social support and happiness scores influences the life expectancy of populations.

The findings from this analysis could help pinpoint areas where interventions would be most impactful, such as strengthening community support systems or implementing policies designed to enhance societal happiness.

By examining these critical factors, this analysis provides valuable insights into the determinants of life expectancy. The results have practical implications for improving global health outcomes and can contribute meaningfully to both academic research and the development of effective policy frameworks.

1 DATA LOADING

The data for this analysis was sourced from multiple platforms: To enhance the readability and interpretability of the dataset, the column names were renamed for better understanding.

The shape of the life expectancy dataset was examined to understand its structure and dimensions. A closer inspection using head() and tail() revealed that the first 48 rows consisted of aggregated data for regional groups, such as the Arab World, Caribbean small states, and other country groupings defined by the World Bank. Since this data did not pertain to individual countries, it was deemed irrelevant for the analysis and was removed.

Additionally, any rows containing null values were excluded to ensure the dataset was clean and complete, avoiding potential issues during analysis. This step helped maintain the integrity of the results by eliminating incomplete records.for life expectancy was used to extract the data directly through the API. Happiness and Social Support Data: Sourced from Kaggle, this dataset was uploaded to a GitHub repository for easier access and was subsequently pulled into the project using its direct URL. This multi-source approach ensured the availability of comprehensive and reliable data for the study.

[6]:

```
import pandas as pd
import pandas_datareader
import warnings
```

```
warnings.simplefilter(action='ignore', category=FutureWarning)

# Check if pandas version is 0.23
if pd.__version__.startswith('0.23'):
    core.common.is_list_like = api.types.is_list_like

from pandas_datareader.wb import download

# Set variables
YEAR = 2019
LE_INDICATOR = 'SP.DYN.LE00.IN'
happiness_url = 'https://raw.githubusercontent.com/ilakkiya-v/project-sen/refs/
 ↪heads/main/happiness.csv'

# Download data using pandas_datareader and reset the index
life_data = download(indicator = LE_INDICATOR, country = 'all', start = YEAR,␣
 ↪end = YEAR).reset_index()
happiness_data = pd.read_csv(happiness_url)
```

[7]:
```
life_data.head()
```

[7]:
```
                          country  year  SP.DYN.LE00.IN
0     Africa Eastern and Southern  2019       63.754752
1      Africa Western and Central  2019       57.500295
2                      Arab World  2019       71.688418
3          Caribbean small states  2019       72.359231
4  Central Europe and the Baltics  2019       77.265533
```

[8]:
```
happiness_data.head()
```

[8]:
```
       country  happiness  Social support
0  Afghanistan      3.203           0.517
1      Albania      4.719           0.848
2      Algeria      5.211           1.160
3    Argentina      6.086           1.432
4      Armenia      4.559           1.055
```

2 DATA CLEANING

To enhance the readability and interpretability of the dataset, the column names were renamed for better understanding.

The shape of the life expectancy dataset was examined to understand its structure and dimensions. A closer inspection using head() and tail() revealed that the first 48 rows consisted of aggregated data for regional groups, such as the Arab World, Caribbean small states, and other country groupings defined by the World Bank. Since this data did not pertain to individual countries, it was deemed irrelevant for the analysis and was removed.

Additionally, any rows containing null values were excluded to ensure the dataset was clean and complete, avoiding potential issues during analysis. This step helped maintain the integrity of the results by eliminating incomplete records.

[10]:

```
print("shape of population table: ",life_data.shape)
print("shape of literacy table: ",happiness_data.shape)
```

```
shape of population table:  (266, 3)
shape of literacy table:  (156, 3)
```

[11]:

```
life_data.rename(columns={'SP.DYN.LE00.IN': 'life'}, inplace=True)
life_data.drop('year', axis=1, inplace=True)
print("Columns of population table:", life_data.columns)
print("Columns of literacy table:", happiness_data.columns)
```

```
Columns of population table: Index(['country', 'life'], dtype='object')
Columns of literacy table: Index(['country', 'happiness', 'Social support'],
dtype='object')
```

[12]:

```
life_data.head(50)
```

[12]:

```
                                               country       life
0                          Africa Eastern and Southern  63.754752
1                           Africa Western and Central  57.500295
2                                           Arab World  71.688418
3                               Caribbean small states  72.359231
4                       Central Europe and the Baltics  77.265533
5                           Early-demographic dividend  70.985650
6                                  East Asia & Pacific  76.787310
7           East Asia & Pacific (excluding high income)  75.992673
8            East Asia & Pacific (IDA & IBRD countries)  76.027483
9                                            Euro area  82.283096
10                               Europe & Central Asia  78.168801
11        Europe & Central Asia (excluding high income)  74.313209
12         Europe & Central Asia (IDA & IBRD countries)  74.322790
13                                      European Union  81.315597
14                Fragile and conflict affected situations  62.230040
15              Heavily indebted poor countries (HIPC)  63.318106
16                                         High income  80.139535
17                                           IBRD only  74.120581
18                                     IDA & IBRD total  71.481364
19                                           IDA blend  61.319259
20                                            IDA only  65.605373
21                                           IDA total  64.169863
22                            Late-demographic dividend  76.990843
23                            Latin America & Caribbean  75.035789
24  Latin America & Caribbean (excluding high income)  74.956870
25  Latin America & the Caribbean (IDA & IBRD coun…  74.970151
26         Least developed countries: UN classification  65.126575
```

```
27                               Low & middle income  71.368288
28                                        Low income  63.434612
29                               Lower middle income  68.740455
30                        Middle East & North Africa  73.845005
31  Middle East & North Africa (excluding high inc…  73.072242
32  Middle East & North Africa (IDA & IBRD countries)  73.047285
33                                     Middle income  72.287301
34                                     North America  79.141358
35                                    Not classified        NaN
36                                      OECD members  80.221309
37                                Other small states  73.454467
38                       Pacific island small states  69.316865
39                         Post-demographic dividend  80.894772
40                          Pre-demographic dividend  60.674126
41                                      Small states  72.734973
42                                        South Asia  70.458293
43                           South Asia (IDA & IBRD)  70.458293
44                                Sub-Saharan Africa  61.211033
45        Sub-Saharan Africa (excluding high income)  61.209933
46         Sub-Saharan Africa (IDA & IBRD countries)  61.211033
47                               Upper middle income  76.025102
48                                             World  72.931034
49                                       Afghanistan  63.565000
```

[13]:

```python
#removing the first 49 values and removing NA values from the rest of the data
life_data = life_data[49:].dropna()
happiness_data = happiness_data.dropna()

print("Shape of life expectancy table: ",life_data.shape)
print("Shape of happiness score table: ",happiness_data.shape)
```

```
Shape of life expectancy table:  (209, 2)
Shape of happiness score table:  (156, 3)
```

[14]:

```python
life_data.head()
```

[14]:

```
                country    life
49          Afghanistan  63.565
50              Albania  79.282
51              Algeria  76.474
54               Angola  62.448
55  Antigua and Barbuda  78.691
```

3 DATA WRANGLING

To prepare the dataset for analysis and visualization, the data sources were merged to create a comprehensive and unified dataset. This step ensured that all relevant variables were consolidated for accurate comparisons and insights.

The life column, which represents life expectancy, was converted into a numeric format to facilitate further processing. This conversion allowed for the creation of categorical groupings based on life expectancy, which were used to enhance the visualization and interpretation of the data. This approach provided a clearer understanding of patterns and trends in the dataset.

[16]:

```
# Merge the result with Happiness data on 'country'
merged_data = pd.merge(life_data, happiness_data, on='country')

merged_data.head()
```

[16]:

```
       country    life  happiness  Social support
0  Afghanistan  63.565      3.203           0.517
1      Albania  79.282      4.719           0.848
2      Algeria  76.474      5.211           1.160
3    Argentina  77.284      6.086           1.432
4      Armenia  75.439      4.559           1.055
```

[17]:

```
merged_data.shape
```

[17]:

```
(133, 4)
```

[18]:

```
merged_data['life'] = pd.to_numeric(merged_data['life'], errors='coerce')

print(merged_data.head())

# Create categorical groupings for Healthy Life Expectancy
merged_data['life_cat'] = pd.cut(merged_data['life'], bins=3,labels=['Low',␣
  ↪'Medium', 'High'])
```

```
       country    life  happiness  Social support
0  Afghanistan  63.565      3.203           0.517
1      Albania  79.282      4.719           0.848
2      Algeria  76.474      5.211           1.160
3    Argentina  77.284      6.086           1.432
4      Armenia  75.439      4.559           1.055
```

[19]:

```
merged_data.to_csv('final-wrangled.csv', index=False)
```

4 DATA ANALYSIS

To assess the relationships between the variables, the correlation factors were calculated for each pair of factors. The analysis revealed strong correlations:

The correlation between life expectancy and happiness score, Between life expectancy and social support, and Between happiness score and social support, all exceeded a value of 0.7. These strong positive correlations validate the focus of the study and indicate that the selected variables are closely related.

Next, the analysis examined the top 10 and bottom 10 countries based on happiness scores.

It is noteworthy that all the top 10 countries with a high happiness index also exhibit high life expectancy, without any exceptions. Conversely, none of the bottom 10 countries with low happiness scores have high life expectancy. All these countries fall into the low or medium-low life expectancy categories, reinforcing the observed relationships between happiness and longevity.

[21]:

```
from scipy.stats import pearsonr


correlation, p_value = pearsonr(merged_data['life'], merged_data['happiness'])
print(f"Correlation between Happiness Score and life expectancy: {correlation:.
 ↪3f}")

correlation, p_value = pearsonr(merged_data['life'], merged_data['Social␣
 ↪support'])
print(f"Correlation between Social Support and life expectancy: {correlation:.
 ↪3f}")

correlation, p_value = pearsonr(merged_data['happiness'], merged_data['Social␣
 ↪support'])
print(f"Correlation between Social Support and Happiness Score: {correlation:.
 ↪3f}")

# Determine the top 10 countries with the highest happiness scores
top_10_happiness = merged_data.sort_values(by='happiness', ascending=False).
 ↪head(10)
print("\nTop 10 Countries with the Highest Happiness Scores:")
print(top_10_happiness.head(10))

# Determine the bottom 10 countries with the highest happiness scores
top_10_happiness = merged_data.sort_values(by='happiness', ascending=True).
 ↪head(10)
print("\nTop 10 Countries with the Highest Happiness Scores:")
print(top_10_happiness.head(10))
```

```
Correlation between Happiness Score and life expectancy: 0.797
Correlation between Social Support and life expectancy: 0.731
Correlation between Social Support and Happiness Score: 0.791

Top 10 Countries with the Highest Happiness Scores:
         country       life  happiness  Social support life_cat
39       Finland  81.982927      7.769           1.587     High
33       Denmark  81.451220      7.600           1.573     High
94        Norway  82.958537      7.554           1.582     High
51       Iceland  83.163415      7.494           1.624     High
88   Netherlands  82.112195      7.488           1.522     High
117  Switzerland  83.904878      7.480           1.526     High
116       Sweden  83.109756      7.343           1.487     High
89   New Zealand  82.056098      7.307           1.557     High
```

```
23        Canada  82.223902      7.278           1.505     High
6        Austria  81.895122      7.246           1.475     High

Top 10 Countries with the Highest Happiness Scores:
                      country    life  happiness  Social support life_cat
113               South Sudan  55.912      2.853           0.575      Low
24   Central African Republic  55.025      3.083           0.000      Low
0                 Afghanistan  63.565      3.203           0.517   Medium
119                  Tanzania  66.989      3.231           0.885   Medium
104                    Rwanda  66.437      3.334           0.711   Medium
73                     Malawi  64.119      3.410           0.560   Medium
16                   Botswana  65.464      3.488           1.145   Medium
48                      Haiti  64.255      3.597           0.688   Medium
132                  Zimbabwe  61.292      3.663           1.114      Low
20                    Burundi  62.351      3.775           0.447      Low
```

5 DATA VISUALISATION

Here we plot 3 digrams:

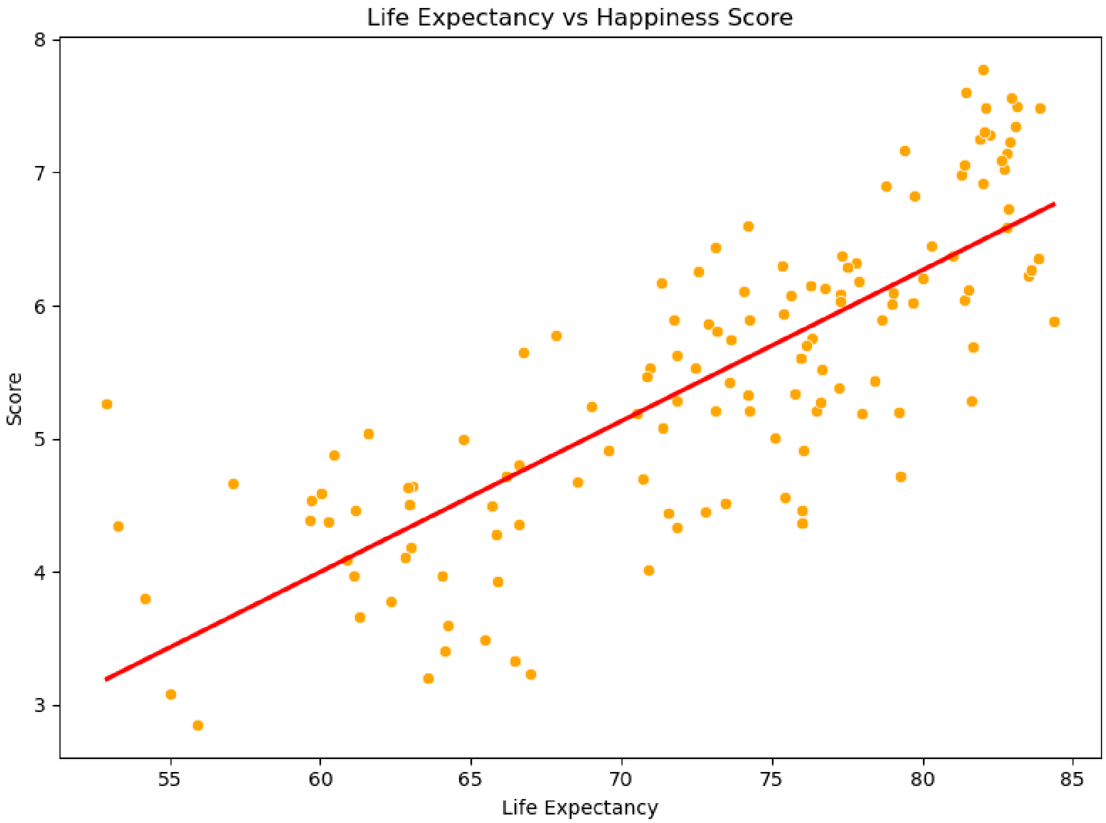
1. ***LINEAR REGRESSION MODEL (scatterplot)***: As life expectancy increases, the happiness score also tends to increase. This suggests that countries with higher life expectancy generally report higher levels of happiness.
2. ***Boxplot***: Countries with 'High' life expectancy have a higher median happiness score compared to 'Medium' and 'Low' categories.
3. ***3D Scatterplot***: The points cluster towards higher values of all three variables, indicating that countries with higher life expectancy and happiness scores also tend to have stronger social support systems.

[23]:

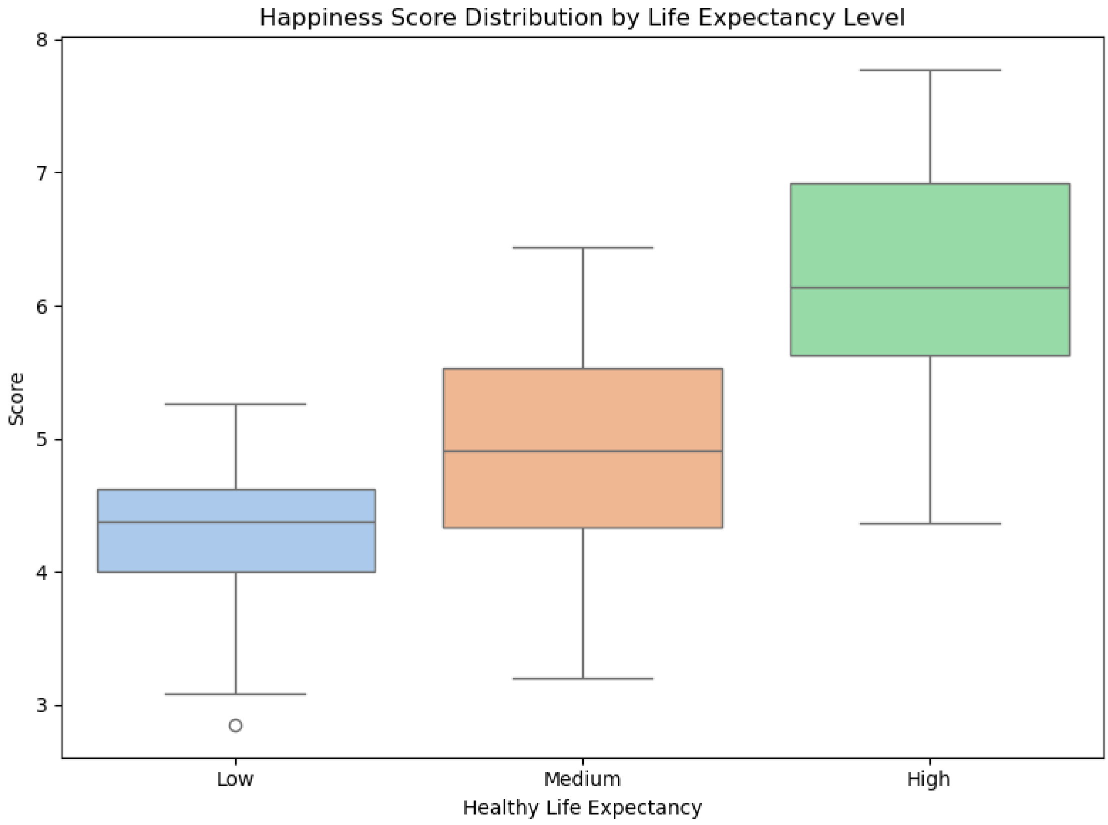
```python
import seaborn as sns
import matplotlib.pyplot as plt

# Scatter plot with linear regression line
plt.figure(figsize=(8, 6))
sns.scatterplot(x='life', y='happiness', data=merged_data, color='orange')
sns.regplot(x='life', y='happiness', data=merged_data, scatter=False,␣
 ↪color='red', ci=None)

# Plot customization
plt.title("Life Expectancy vs Happiness Score")
plt.xlabel("Life Expectancy")
plt.ylabel("Score")
plt.tight_layout()
plt.show()
```

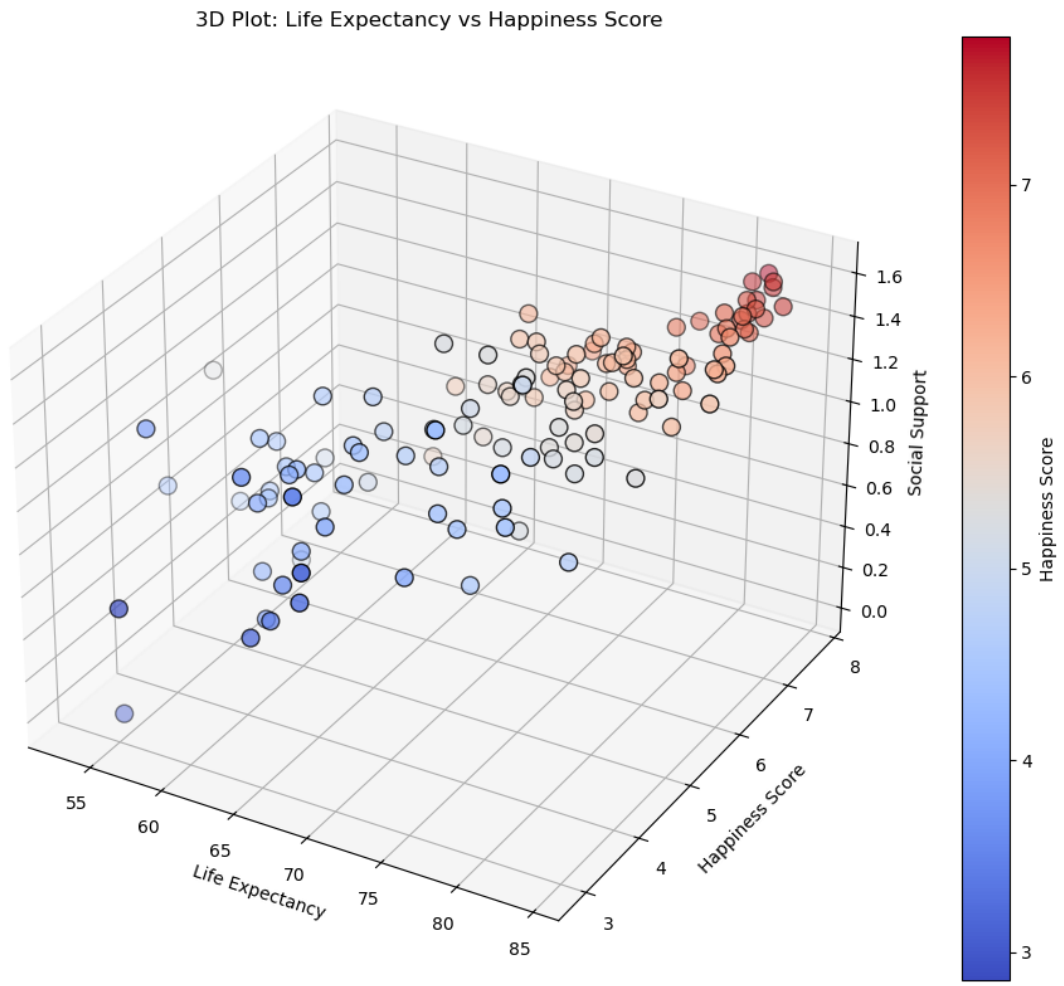


```
[24]: # Box plot of Happiness Scores by Life Expectancy Group
      plt.figure(figsize=(8, 6))
      sns.boxplot(x='life_cat', y='happiness', data=merged_data, palette='pastel')
      plt.title("Happiness Score Distribution by Life Expectancy Level")
      plt.xlabel("Healthy Life Expectancy")
      plt.ylabel("Score")
      plt.tight_layout()
      plt.show()
```



[25]:

```
fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111, projection='3d')
scatter = ax.scatter(merged_data['life'], merged_data['happiness'],␣
  ↪merged_data['Social support'],
                     c=merged_data['happiness'], cmap='coolwarm', s=100,␣
  ↪edgecolor='k')
ax.set_title("3D Plot: Life Expectancy vs Happiness Score")
ax.set_xlabel("Life Expectancy")
ax.set_ylabel("Happiness Score")
ax.set_zlabel("Social Support")
fig.colorbar(scatter, ax=ax, label='Happiness Score')
plt.tight_layout()
plt.show()
```



Future Improvements: 1. Expand the analysis to include multiple years for a longitudinal perspective.

2. Incorporate additional variables like healthcare access and education levels.

GITHUB LINK

https://github.com/senthil1814/2507127_MN5813

DATA SOURCES: * Happiness score and social support: https://www.kaggle.com/datasets/unsdsn/world-happiness/data?select=2019.csv

- Life expectancy : https://data.worldbank.org/indicator/SP.DYN.LE00.IN