World Happiness Data Exploration, Visualization and Pre-processing Report

Sandrine Mutoni

Dennis Zittel

Project Mentor: Tarik Anouar

Table of Contents

1. Introduction

This report presents an exploratory data analysis of the World Happiness dataset compiled by the United Nations. The project aims to understand which factors most strongly influence national happiness and to visualize global and regional trends across the years. The dataset includes information such as GDP, social support, life expectancy, freedom of choice, and emotional responses like joy and sadness.

All analyses were conducted in Google Colab using Python libraries such as Pandas, Seaborn, and Plotly. The project covers the period from 2005 to 2021 and involves merging multiple datasets, integrating regional metadata, and performing statistical checks to prepare a robust and consistent final dataset.

2. Data Sources

The project makes use of the following three datasets:

1. **World Happiness Report (2005–2020)**: Provides global happiness indicators over 16 years in consistent format.

https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021?resource=download&select=world-happiness-report.csv

2. **World Happiness Report 2021**: Contains updated country data but in a slightly different structure, requiring alignment.

https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021?resource=download&select=world-happiness-report-2021.csv

3. **Continents Classification (OWID)**: A reference dataset mapping each country to its global region (Africa, Europe, etc.).

https://ourworldindata.org/world-region-map-definitions

These files were uploaded to a dedicated Google Drive folder and accessed from within the Google Colab environment.

3. Data Preparation & Cleaning

This section describes all steps taken to prepare the raw data for analysis. These steps were carried out in Google Colab and include importing, inspecting, cleaning, transforming, and merging datasets. Special care was taken to standardize column names, unify the time structure, and add useful metadata like regional classification.

3.1 Loading the Environment and Data Files

The first step involved connecting Google Colab to Google Drive. This allowed access to the project folder that contains all necessary CSV files. A specific path was defined to access the files. This setup enabled the notebook to read files directly from the structured folder system in Drive.

3.2 Importing the Historical Dataset (2005–2020)

The historical dataset was imported using the Pandas library. Initially, the dataset was inspected to check its structure and contents. This involved displaying the first few rows to get an overview of the data, reviewing the general structure of the dataset, and analyzing basic statistics.

The dataset includes the following key variables:

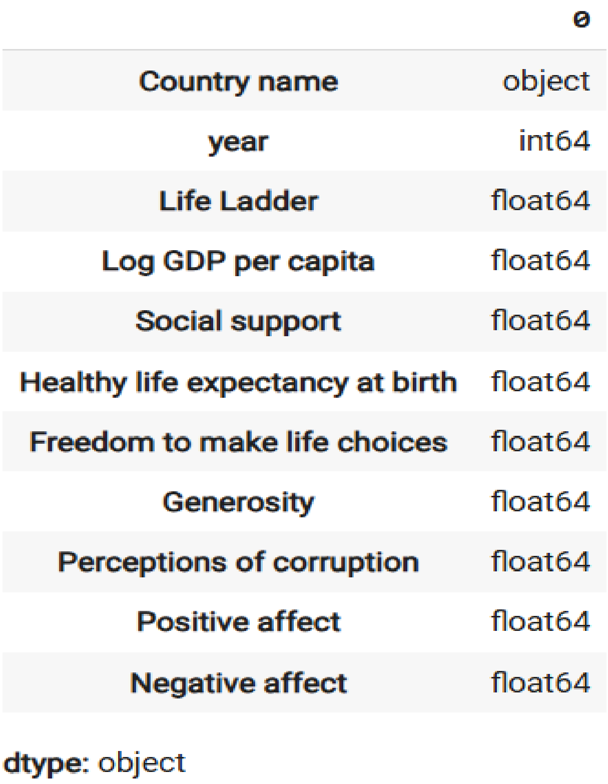
- **Country name**: The name of the country or region, indicating which country the data is associated with.
- **Life Ladder**: This represents the overall happiness score for each country, based on respondents' self-reported levels of happiness.
- **Log GDP per capita**: The logarithm of the Gross Domestic Product per capita, representing the economic output per person in a country.
- **Social support**: This variable indicates the level of support that individuals in a country feel they can rely on, such as help from friends or family in times of need.
- **Healthy life expectancy at birth**: This represents the expected number of years a newborn is expected to live in full health.
- **Freedom to make life choices**: This variable measures the level of personal freedom individuals feel they have in making choices in their lives.
- **Generosity**: This is a measure of the willingness of individuals in a country to donate time or money to charitable causes.

- **Perceptions of corruption**: This represents how people in each country perceive corruption in their government and business sectors.
- **Positive affect**: This variable captures the frequency of positive emotions such as happiness, enjoyment, or laughter experienced by individuals in each country.
- **Negative affect**: This measures the frequency of negative emotions such as sadness, anger, or worry experienced by individuals.

The dataset confirmed that both numeric and categorical variables were included, which are essential for performing the analysis in the following sections.

3.3 Analyzing Data Types, Missing Values and Categorizing Variables

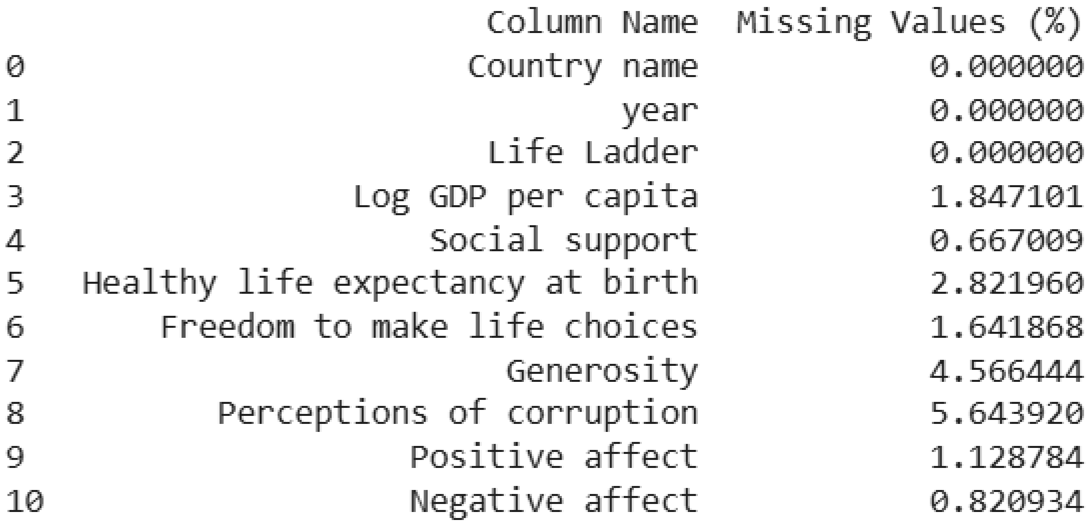
First, the data types of all columns were analyzed using df.dtypes. This provided a clear understanding of which columns were numerical and which were of object (categorical) type.



	0
Country name	object
year	int64
Life Ladder	float64
Log GDP per capita	float64
Social support	float64
Healthy life expectancy at birth	float64
Freedom to make life choices	float64
Generosity	float64
Perceptions of corruption	float64
Positive affect	float64
Negative affect	float64

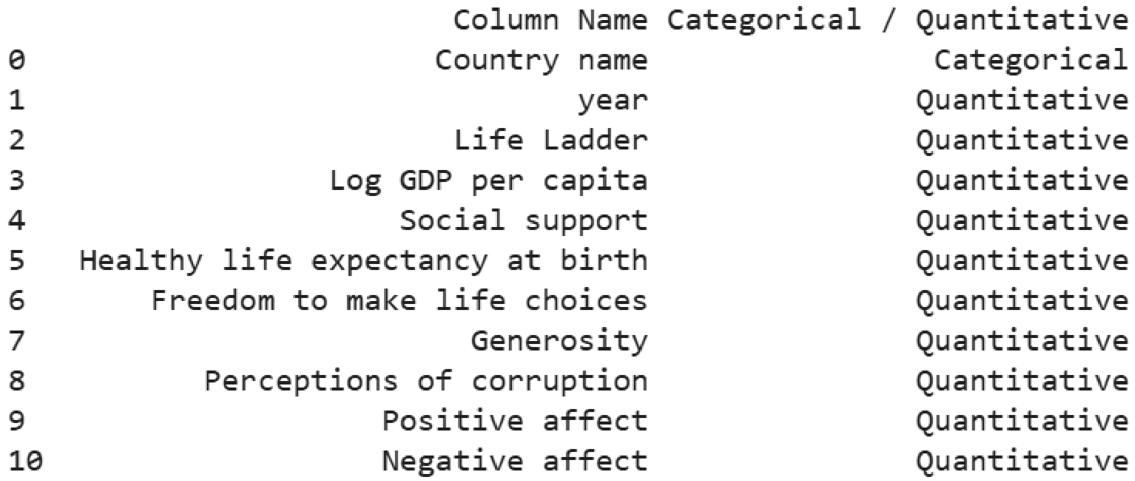
dtype: object

Next, missing values were, to express the extent of missing data as a percentage of the total number of rows, the following calculation was performed. The Table below presents the results for each variable.



```
                       Column Name  Missing Values (%)
0                     Country name            0.000000
1                             year            0.000000
2                      Life Ladder            0.000000
3               Log GDP per capita            1.847101
4                   Social support            0.667009
5   Healthy life expectancy at birth          2.821960
6        Freedom to make life choices         1.641868
7                       Generosity            4.566444
8         Perceptions of corruption           5.643920
9                   Positive affect           1.128784
10                  Negative affect           0.820934
```

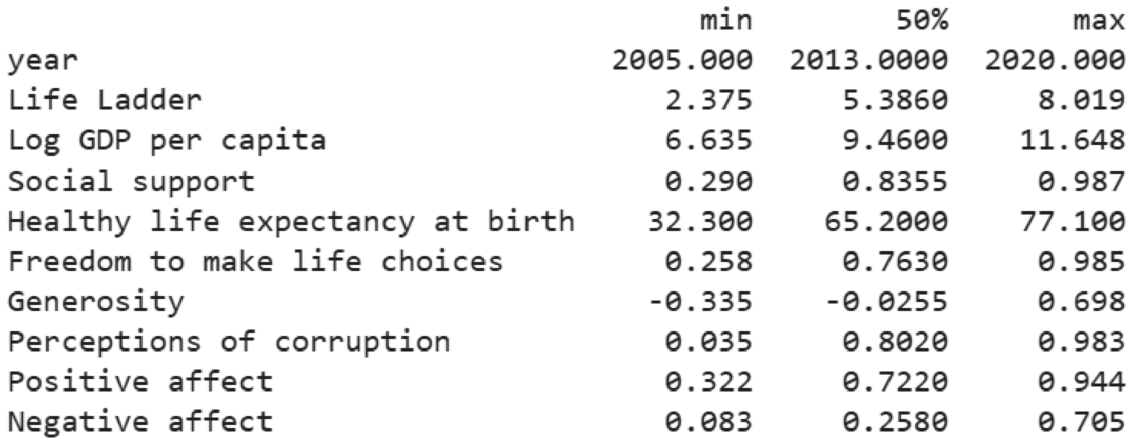
Finally, each column was categorized as either categorical or quantitative based on its data type using a custom classification loop. The number of unique values in each categorical column was also calculated to support deeper analysis and visualization planning. The Table below presents the results for each variable.



```
                         Column Name Categorical / Quantitative
0                       Country name                Categorical
1                               year               Quantitative
2                        Life Ladder               Quantitative
3                 Log GDP per capita               Quantitative
4                     Social support               Quantitative
5   Healthy life expectancy at birth               Quantitative
6       Freedom to make life choices               Quantitative
7                         Generosity               Quantitative
8          Perceptions of corruption               Quantitative
9                    Positive affect               Quantitative
10                   Negative affect               Quantitative
```

3.4 Summarizing Distributions

Descriptive statistics for all quantitative columns were reviewed. The Table below presents the results for each variable.



```
                                       min        50%       max
year                              2005.000  2013.0000  2020.000
Life Ladder                          2.375     5.3860     8.019
Log GDP per capita                   6.635     9.4600    11.648
Social support                       0.290     0.8355     0.987
Healthy life expectancy at birth    32.300    65.2000    77.100
Freedom to make life choices         0.258     0.7630     0.985
Generosity                          -0.335    -0.0255     0.698
Perceptions of corruption            0.035     0.8020     0.983
Positive affect                      0.322     0.7220     0.944
Negative affect                      0.083     0.2580     0.705
```

This provided insights into the range of values and potential outliers across features like GDP, life expectancy, and corruption perception.

3.5 Preparing the 2021 Dataset

The 2021 dataset was imported with Pandas and stored in df_2021_raw. Since it had different column names, the necessary variables were renamed using df.rename() to align with the older dataset. The following key variables were kept:

- Country name
- Life Ladder
- Log GDP per capita
- Social support
- Healthy life expectancy at birth
- Freedom to make life choices
- Generosity
- Perceptions of corruption

A new column year = 2021 was manually added.

3.6 Merging the Datasets

The historical dataset was renamed as df_old, and both datasets were concatenated into one.

This resulted in a combined dataset containing all years from 2005 to 2021. The use of ignore_index=True ensured a clean, continuous index.

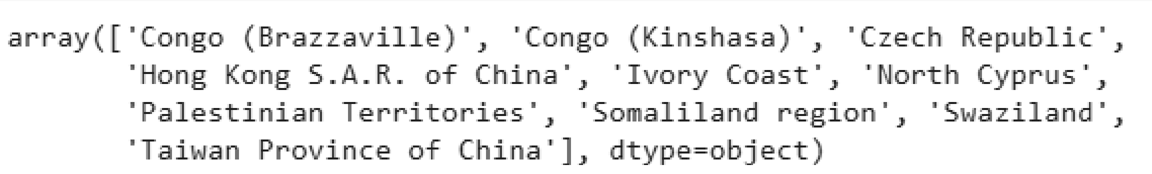
3.7 Region Mapping and Final Dataset Overview

To enrich the dataset with geographical context, a third dataset from Our World in Data (OWID) (https://ourworldindata.org/world-region-map-definitions) was used to assign each country to a world region (such as Europe, Asia, Africa, etc.). This process involved several transformation steps.

The OWID CSV file contained more columns than needed. Therefore, we renamed and filtered only the needed columns for the merge.

The regional information was then merged with the main DataFrame df_merged using a left join on "Country name" to preserve all existing records. This created the final working DataFrame df_final, which now contained 12 columns, including the newly added "Region" field.

After the merge, a check was performed to identify which countries still had missing region assignments (NaN values).



```
array(['Congo (Brazzaville)', 'Congo (Kinshasa)', 'Czech Republic',
       'Hong Kong S.A.R. of China', 'Ivory Coast', 'North Cyprus',
       'Palestinian Territories', 'Somaliland region', 'Swaziland',
       'Taiwan Province of China'], dtype=object)
```

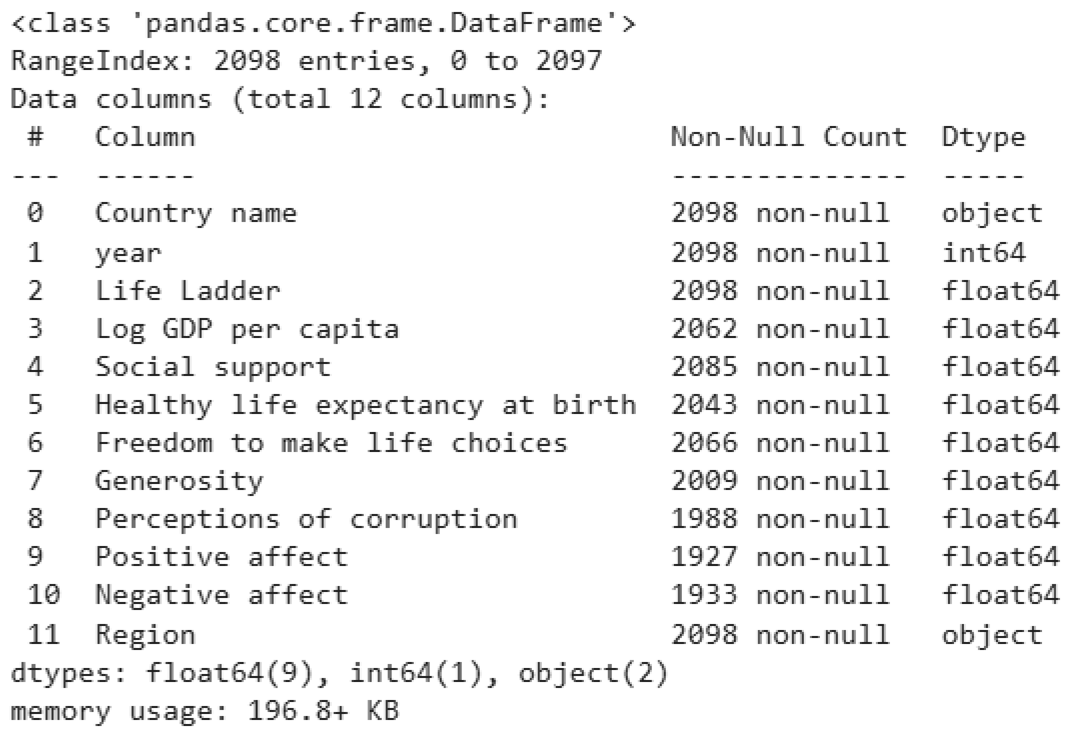
Because these countries had naming variations or were not listed in the OWID dataset, manual corrections were applied as follows:

These assignments ensured full regional coverage across the dataset and eliminated all NaN values in the "Region" column.

After merging and manual cleaning, the resulting dataset df_final contains:

- 2098 rows
- 12 clean and analysis-ready columns, including the "Region" field.

The dataset is now fully enriched with geographic metadata and ready for further exploratory data analysis and visualization.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2098 entries, 0 to 2097
Data columns (total 12 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   Country name                      2098 non-null   object 
 1   year                              2098 non-null   int64  
 2   Life Ladder                       2098 non-null   float64
 3   Log GDP per capita                2062 non-null   float64
 4   Social support                    2085 non-null   float64
 5   Healthy life expectancy at birth  2043 non-null   float64
 6   Freedom to make life choices      2066 non-null   float64
 7   Generosity                        2009 non-null   float64
 8   Perceptions of corruption         1988 non-null   float64
 9   Positive affect                   1927 non-null   float64
 10  Negative affect                   1933 non-null   float64
 11  Region                            2098 non-null   object 
dtypes: float64(9), int64(1), object(2)
memory usage: 196.8+ KB
```

4. Exploratory Data Visualization

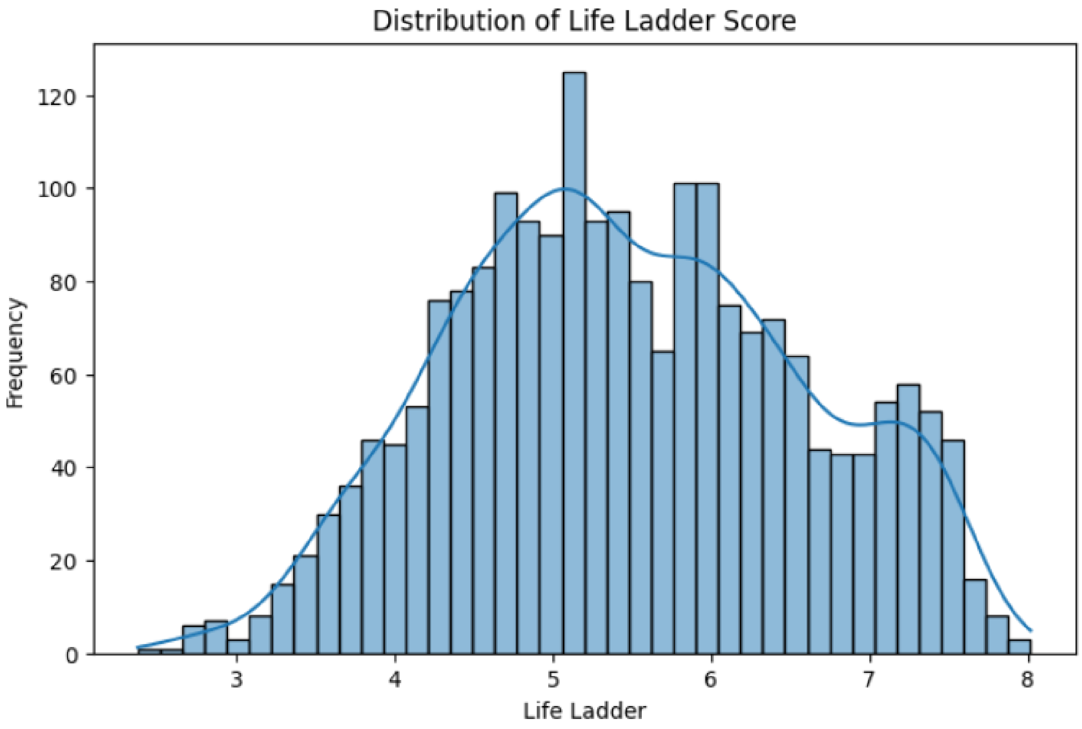
Introduction

This section provides key visual insights from the World Happiness Report. We explore the distribution of Ladder scores, analyze relationships using a correlation Heat-map, and highlight the top happiest and unhappiest Countries. A global map shows happiness by location, while a scatter plot examines the link between GDP and Life Ladder . Finally , a line chart illustrates how happiness score have changed across regions from 2005 to 2021.

4.1 Distribution of Life Ladder Score

Description

This histogram plot represents the distribution of Life Ladder scores (a measure of happiness) from the World Happiness dataset. The x-axis represents the Life Ladder scores, while the y-axis represents the frequency (count of occurrences) of these scores in the dataset.



Observations

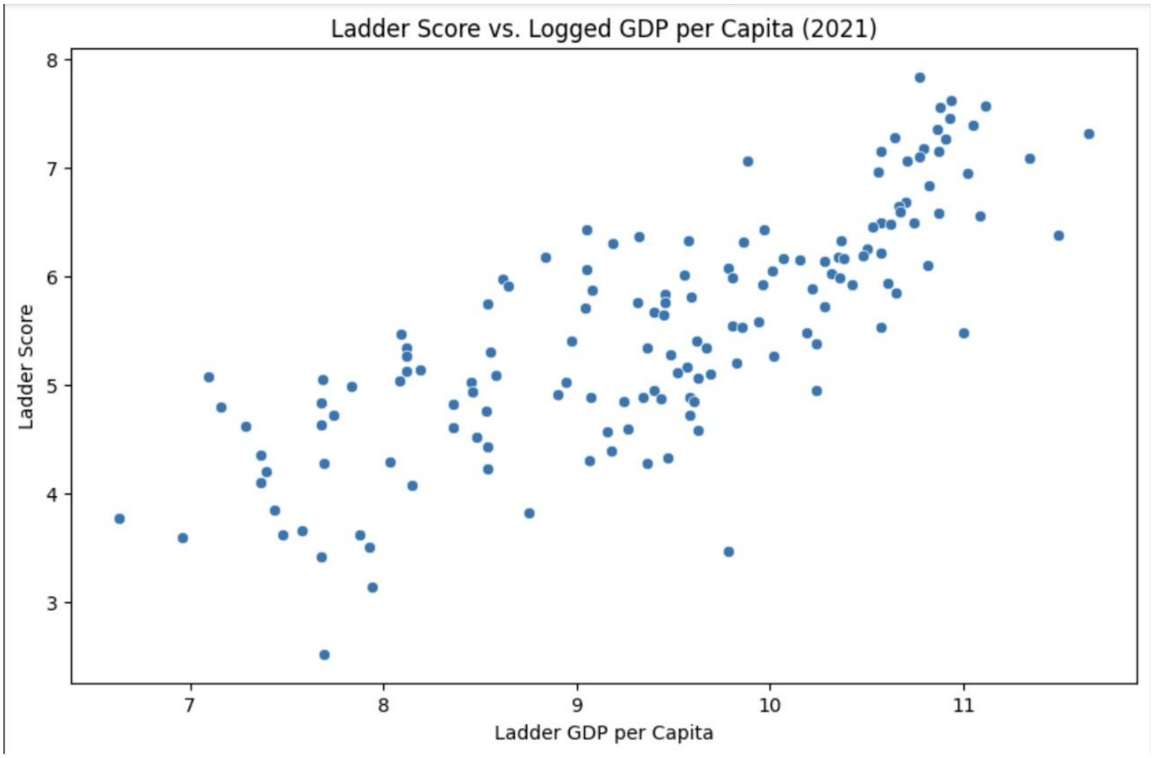
Normal distribution and most countries have happiness scores that cluster around the middle range, with fewer countries at the extreme low or high ends.

The highest frequency of scores falls between 5.5 and 6.5, and many countries report moderate levels of happiness.

The scores range from around 2 to 8, meaning that some countries have very low happiness while others enjoy significantly high levels.

Commentary

Countries with higher happiness scores likely have strong economies, good healthcare, and social support, while lower scoring countries may struggle with economic instability or Political issues.

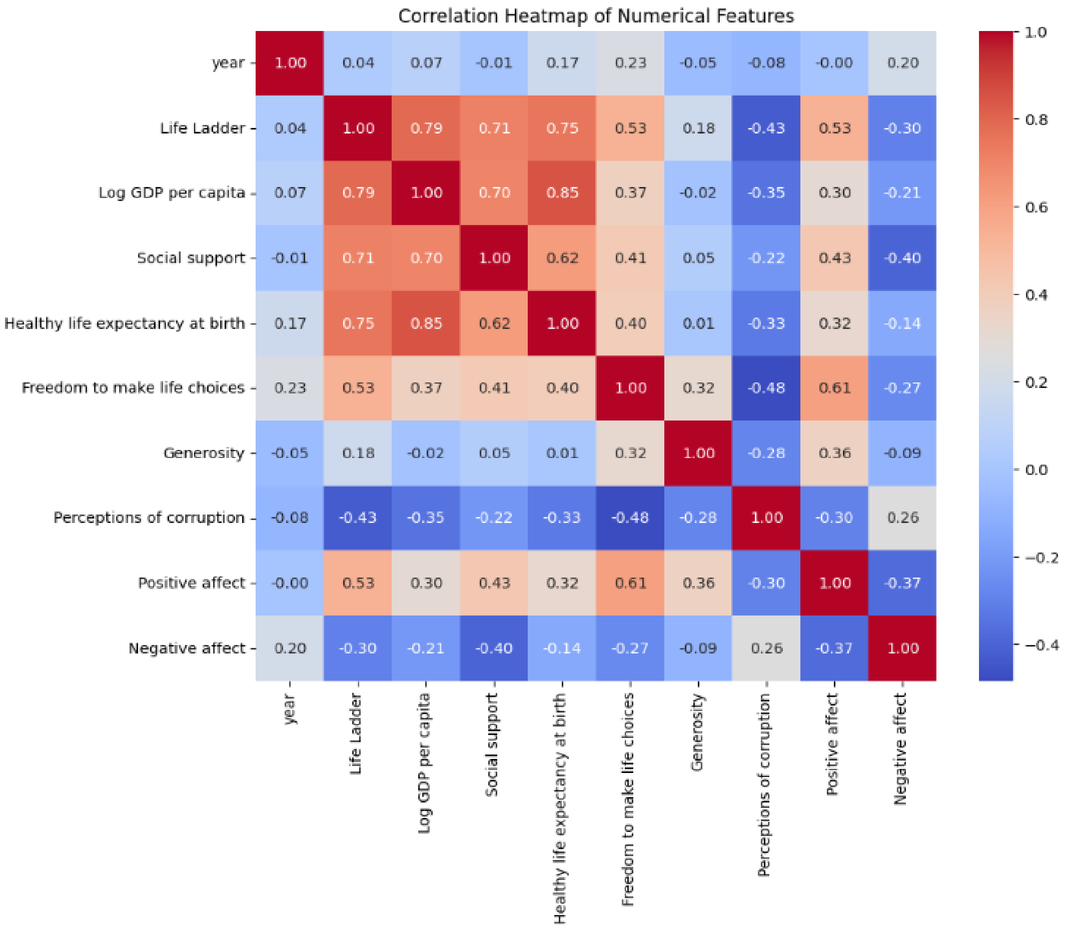


This scatter plot illustrates the relationship between a country’s economic wealth and its people’s life satisfaction in 2021. We see a strong positive correlation countries with higher income levels tend to report higher income levels tend to report higher happiness scores.

4.2 The correlation Heat map of Numerical Feature

Description

The correlation heatmap visually represents the relationships between numerical features in the World Happiness dataset. Each cell shows the correlation coefficient between two variables, ranging from -1(strong negative correlation) to 1(strong positive correlation). Red indicates positive correlations, while blue indicates negative correlations.



Observations

Strong Positive Correlations: Life Ladder is highly correlated with Log GDP per capita (0.79), Social Support (0.71), and Healthy life expectancy (0.75).

Log GDP per capita also has strong correlations with Social Support (0.70) and Healthy life expectancy (0.85).

Negative Correlations: Perceptions of corruption has a negative correlation with Life Ladder (-0.43) and Social Support (-0.40), meaning higher corruption perception is linked to lower happiness and social support.

Generosity does not show strong correlation with happiness or other economic indicators.

Freedom to make life choices has a moderate correlation with Life Ladder (0.53) and Positive affect (0.61), that people who feel free in their choices tend to be happier.

Commentary

Corruption perception negatively impacts happiness, likely due to a lack of trust in institutions.

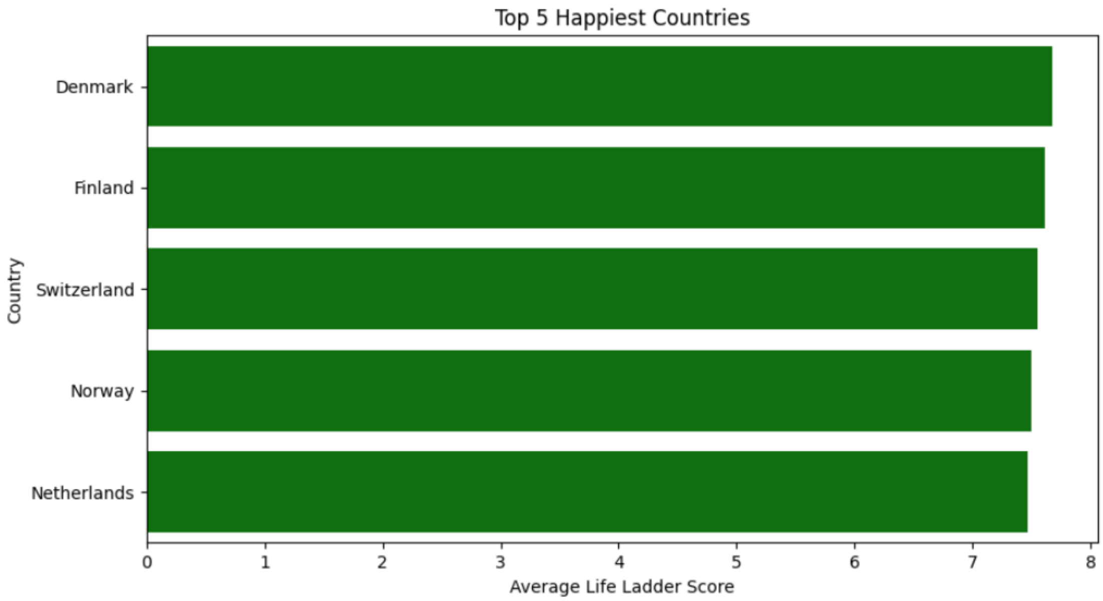
Generosity shows weak correlations, indicating that economic and social factors might play a bigger role in overall happiness than individual acts of giving.

Policies focusing on economic growth, healthcare, and social trust could lead to higher happiness scores in different countries.

4.3 Top Happiest Countries and Unhappiest Countries

Description

The x-axis represents the Average Life Ladder Score, a metric used to quantify happiness levels, The y-axis lists the top five happiest countries.



Observations

All five countries have high Life Ladder scores, suggesting a very high standard of living, economic stability, social support, and overall well-being.

The score appears close to each other, meaning there isn't a massive gap between the top-ranked countries.

Commentary

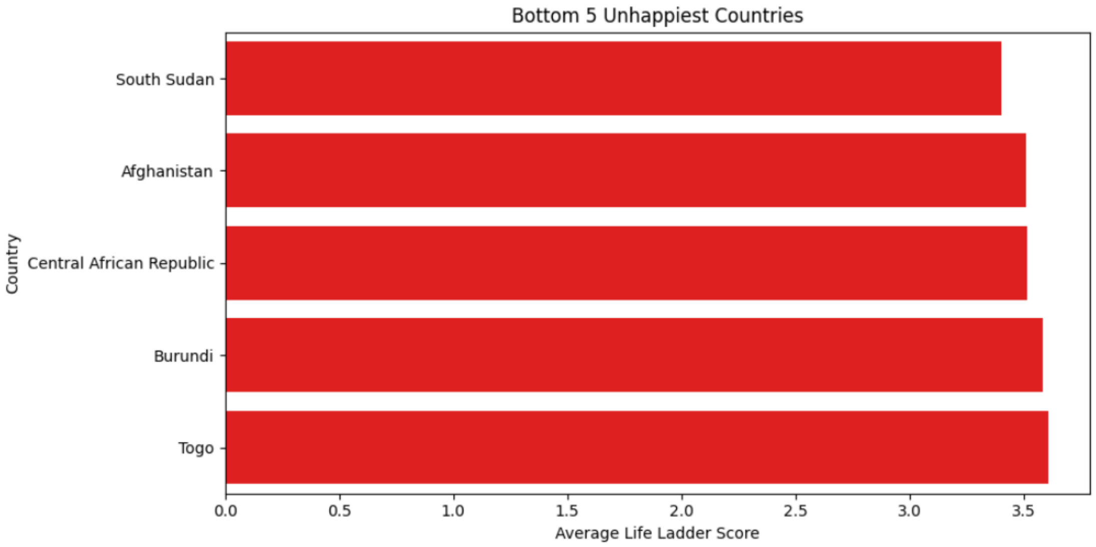
The results align with global happiness research, which often ranks Nordic and European nations among the happiest due to factors like strong social safety nets, excellent healthcare, high GDP per capita, work-life balance, and high levels of trust in government.

Denmark, Finland, and Switzerland regularly rank at the tops in annual happiness reports.

Unhappiest Country

Description

The x-axis represents the Average Life Ladder Score, a metric used to quantify happiness levels, The y-axis lists the top five unhappiest countries.



Observations

South Sudan appears to be the unhappiest country, having the lowest score among the five.

The other four countries—Afghanistan, Central African Republic, Burundi, and Togo—also have very low happiness levels, suggesting similar socio-economic struggles.

The ranking aligns with global trends where war, poverty, political instability, and poor economic conditions contribute to low happiness scores.

Commentary

The unhappiness in South Sudan could be attributed to ongoing conflicts, political instability, and economic struggles, including food insecurity and displacement of people.

Afghanistan has faced continued instability due to conflicts, regime changes, and humanitarian crises, significantly affecting the well-being of its citizens.

The Central African Republic and Burundi have struggled with political instability, underdevelopment, and economic hardships, which might explain their low happiness scores.

Togo, although showing some economic improvements, still deals with poverty and social inequality, impacting overall happiness.

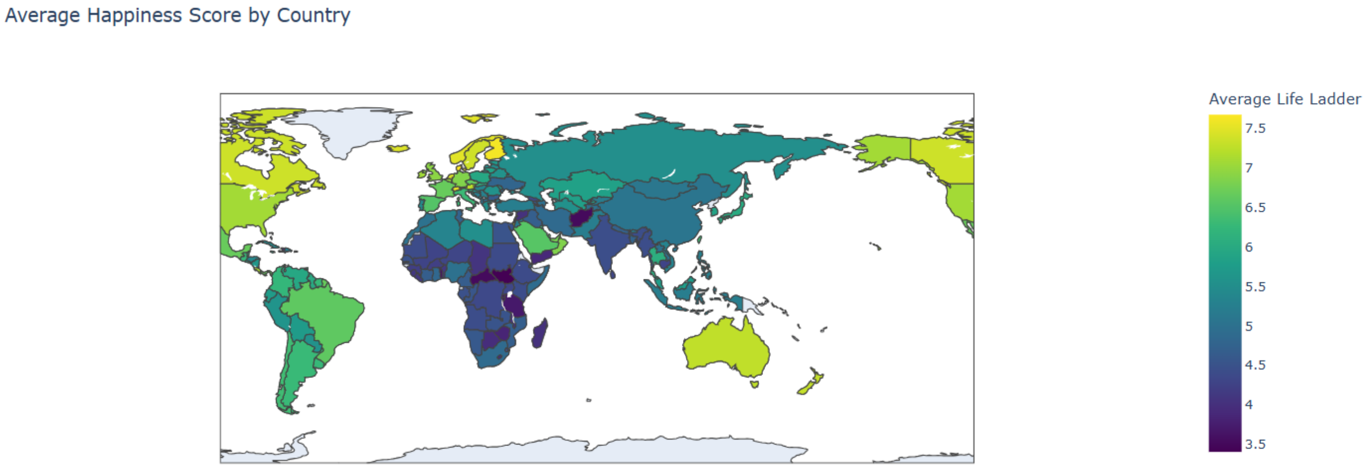
4.4 Global Happiness Map

Description

The color scale on the right indicates happiness levels, with dark blue representing the lowest scores (around 3.5) and bright yellow green representing the highest scores (around 7.5).

The world map is color-coded based on happiness scores (often measured by the World Happiness Report using the Life Ladder Score)

Different regions of the world have varying happiness levels, providing a visual representation of global well-being.



Observation

Countries are color-coded based on happiness levels;

Yellow & Green → High Happiness scores.

Blue & Purple → Lower Happiness scores.

From the map, we can see that Northern Europe, North America, and Australia rank among the happiest places, while Sub-Saharan Africa and parts of Asia show lower scores.

A clear north-south divide is visible, with richer, more developed nations tending to score higher.

There is a strong correlation between economic prosperity, social support, and happiness.

Commentary

Countries like Finland, Denmark, and Switzerland consistently rank highest due to:

- Strong social welfare systems
- High GDP per capita
- Excellent work-life balance

In contrast, many African and conflict-affected countries have lower happiness scores due to:

- Economic struggles
- Limited healthcare
- Political instability

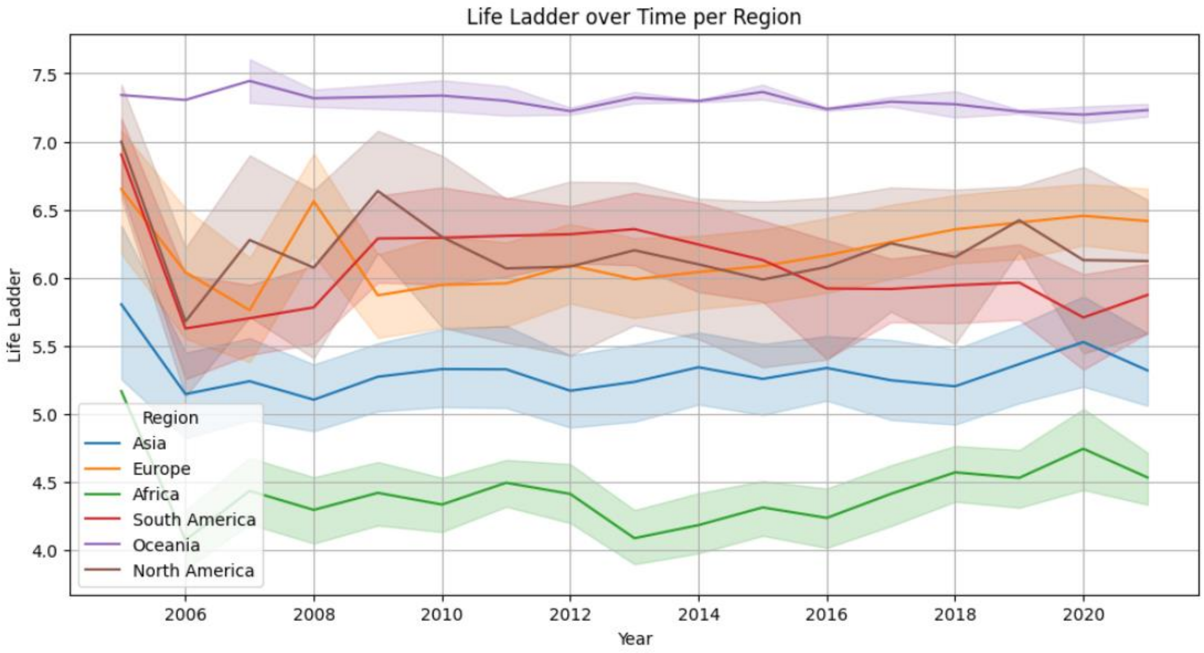
4.5 Life Ladder over Time per Region

Descriptions

This line plot shows how the average happiness scores (Life Ladder) have changed from 2005 to 2021 across different world regions.

Each colored line represents a region (Asia, Europe, Africa, South America, Oceania, North America).

The shaded areas around the lines indicate the variability or uncertainty in the data.



Observations

Oceania (Purple Line) & North America (Dark Red Line) consistently show the highest happiness scores, averaging around 7.0 - 7.5.

Europe (Orange Line) maintains a stable happiness level, staying above 6.0.

Africa (Green Line) has the lowest happiness scores, ranging around 4.0 - 4.5, with some fluctuations.

Asia (Blue Line) remains below the global average, fluctuating around 5.0 - 5.5

Commentary

The happiness gap between regions is clear—wealthier regions (North America, Oceania, Europe) have higher scores, while developing regions (Africa, parts of Asia) struggle.

Economic and social factors like GDP, governance, healthcare, and freedom influence these trends.

Oceania's stability suggests that high life expectancy, good governance, and social support contribute to happiness.

Africa's lower scores may reflect economic hardships, political instability, and healthcare challenges

5. Conclusion

The World Happiness Report analysis provided insightful patterns and relationships across countries and regions from 2005 to 2021. Through careful data cleaning, merging, and enrichment with regional classifications, the dataset was transformed into a robust and analysis-ready resource. Exploratory visualizations revealed strong correlations between happiness and factors such as GDP per capita, social support, and life expectancy, while negative perceptions of corruption were linked to lower happiness levels.

The findings highlight a consistent divide between wealthier and less developed regions, with countries in Northern and Western Europe, North America, and Oceania ranking highest in happiness, while those in conflict-affected or economically challenged areas like Sub-Saharan Africa and parts of Asia showed significantly lower scores.

This project underlines the importance of economic stability, healthcare, freedom, and trust in institutions in fostering national well-being. The insights generated can support policymakers, researchers, and global organizations in designing targeted interventions aimed at improving quality of life and reducing disparities in happiness worldwide.