Exploring Global Happiness: A Comprehensive Analysis of the Impact of Socio-Economic Factors and Health on Happiness Scores Across Regions

Prakruthi Harish

2024-08-19

Section 1: Background

This project addresses the crucial question of what drives global happiness, moving beyond traditional economic measures like GDP. It focuses on how various socio-economic factors influence the Life Ladder score, a key indicator of happiness, and how these influences differ across regions and countries. Understanding these dynamics is vital for policymakers aiming to improve citizens' quality of life. Using data from the World Happiness Report, the project analyzes variables such as GDP per capita, social support, life expectancy, freedom of choice, generosity, and perceptions of corruption.

By exploring these factors, the project highlights patterns and correlations that explain variations in happiness globally, offering data-driven insights for policy interventions. Helliwell et.al (2019) in the World Happiness Report emphasized the impact of economic and social factors on happiness. This project expands on that analysis by incorporating a broader range of economic measures and offering a multi-regional perspective. By examining these factors over time, the project provides deeper insights into the trends that shape global happiness.

The primary objective is to understand the correlation between various factors and their impact on happiness, identifying region-specific trends and global patterns. The project employs exploratory data analysis (EDA), using statistical tests and visualizations like histograms, Q-Q plots, box plots, and interactive dashboards. These findings offer actionable solutions for enhancing well-being worldwide.

Section 2: Methods

This project analyzes two key datasets from the World Happiness Report: one covering 2005-2023 for a broad temporal view, and another focused on 2024 for current socio-economic insights. These datasets, sourced from the World Happiness Report and Kaggle, are vital for understanding both short-term and long-term global well-being. The data, gathered through the Gallup World Poll, includes metrics like GDP per capita, social support, life expectancy, and more, across 150+ countries. The 2024 dataset contains 143 rows and 12 columns, while the 2005-2023 dataset has 2363 rows and 11 columns.

2.2. Data Cleaning and Pre-processing

It is essential to ensure that the data is clean and ready for exploration before starting the analysis. So the pre-processing step involved addressing missing values, non-ASCII characters, and other potential data inconsistencies. Missing values were handled by imputing them with the mean of the respective columns. This approach ensured that the data remained representative and accurate. Additionally, non-ASCII characters were removed from the country names to avoid errors during analysis and visualization. These pre-processing steps were crucial to maintain the integrity of the dataset and to ensure reliable results.

2.3. Exploratory Data Analysis (EDA)

The first step in the analysis carried out in this project was to explore the distribution of key variables, including Ladder score, GDP per capita, and Social Support. This was accomplished by visualizing the data through histograms, which provided initial insights into the overall trends and highlighted any anomalies or patterns in the data. However, while histograms provided a general sense of the data distribution, they did not offer detailed insights into the normality of the data. To address this, Q-Q plots were generated to more assess whether the variables followed a normal distribution. This step was crucial for determining the suitability of parametric statistical tests in subsequent analyses. However, Q-Q plots alone did not reveal how these distributions varied across different regions. In order to fill this gap box plots were employed to examine the regional distribution of Ladder scores within the 2024 dataset, offering a more granular view of how happiness varied across different regions. To further contextualize these findings, a pie chart was created to visualize the distribution of countries by region. This step clarified the sample size for each region, thereby enhancing the interpretation of the box plot findings. While this added a layer of understanding about regional representation, it still didn't explain how factors like health might influence happiness.

Further analysis focused on the relationship between health and happiness, with the average Healthy Life Expectancy across regions being plotted to explore how disparities in health might influence regional happiness levels. An interactive choropleth map was then created to track changes in happiness scores over time. This dynamic approach allowed for a comprehensive understanding of global happiness trends over the years. However, this broader view needed to be complemented by a more focused analysis. Thus, a comparative analysis was conducted to identify the happiest and unhappiest countries based on their cumulative Ladder scores. This helped pinpoint specific countries that consistently ranked at the top or bottom of the happiness spectrum. This step was followed by statistical validation through normality checks and T-tests. The correlations between various socio-economic factors and happiness were analyzed, both at the country and regional levels. To visualize the relationships between the Ladder score and various socio-economic factors, scatter plots with regression lines were created which confirmed the linearity and strength of these relationships. This provided a detailed view, yet to enhance the depth of analysis, an advanced interactive visualization was created. This allowed for a more granular exploration of how different factors influence happiness across regions. To further deepen the analysis, an advanced interactive visualization was created. This allowed for a more granular exploration of how different factors influence happiness across regions, offering a more nuanced understanding of the complex dynamics at play.

The code to implement each of these methods, along with the results obtained, is presented below. The results include a series of visualizations and analyses that provide insights into the key drivers of happiness across the globe.

Note: Please scroll down the Slidey page to explore the results and detailed analysis.

Section 3: Implementation and Results

3.1. Install and load libraries

This step loads all the packages required for exploratory data visualization and analysis conducted in this project.

```
##
## The downloaded binary packages are in
##  /var/folders/9c/v_4gdjk147ndq15wh3wpcpy80000gp/T//RtmpeJiYG6/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/v_4gdjk147ndq15wh3wpcpy80000gp/T//RtmpeJiYG6/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/v_4gdjk147ndq15wh3wpcpy80000gp/T//RtmpeJiYG6/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/v_4gdjk147ndq15wh3wpcpy80000gp/T//RtmpeJiYG6/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/v_4gdjk147ndq15wh3wpcpy80000gp/T//RtmpeJiYG6/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/v_4gdjk147ndq15wh3wpcpy80000gp/T//RtmpeJiYG6/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/v_4gdjk147ndq15wh3wpcpy80000gp/T//RtmpeJiYG6/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/v_4gdjk147ndq15wh3wpcpy80000gp/T//RtmpeJiYG6/downloaded_packages
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
##
## Attaching package: 'kableExtra'
##
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
##
##
## Loading required package: SnowballC
##
## corrplot 0.94 loaded
##
##
## Attaching package: 'plotly'
##
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
##
## The following object is masked from 'package:stats':
##
##     filter
##
##
## The following object is masked from 'package:graphics':
##
##     layout
##
##
##
## Attaching package: 'reshape2'
##
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

3.2. Load the Datasets

This step imports the datasets and provides a preliminary understanding of structure and content of data.This step helps identify data types and certain variable format inconsistencies in the datasets early in the process.

```
## The 2024 dataset has 143 rows and 12 columns.
```

```
## The 2005-2023 dataset has 2363 rows and 11 columns.
```

```
## 'data.frame':    143 obs. of  12 variables:
##  $ Country.name                : chr  "Finland" "Denmark" "Iceland" "Sweden" ...
##  $ Regional.indicator          : chr  "Western Europe" "Western Europe" "Western
Europe" "Western Europe" ...
##  $ Ladder.score                : num  7.74 7.58 7.53 7.34 7.34 ...
##  $ upperwhisker                : num  7.82 7.67 7.62 7.42 7.41 ...
##  $ lowerwhisker                : num  7.67 7.5 7.43 7.27 7.28 ...
##  $ Log.GDP.per.capita          : num  1.84 1.91 1.88 1.88 1.8 ...
##  $ Social.support              : num  1.57 1.52 1.62 1.5 1.51 ...
##  $ Healthy.life.expectancy     : num  0.695 0.699 0.718 0.724 0.74 0.706 0.704 0.708
0.747 0.692 ...
##  $ Freedom.to.make.life.choices: num  0.859 0.823 0.819 0.838 0.641 0.725 0.835
0.801 0.759 0.756 ...
##  $ Generosity                  : num  0.142 0.204 0.258 0.221 0.153 0.247 0.224
0.146 0.173 0.225 ...
##  $ Perceptions.of.corruption   : num  0.546 0.548 0.182 0.524 0.193 0.372 0.484
```

```
0.432 0.498 0.323 ...
##  $ Dystopia...residual        : num  2.08 1.88 2.05 1.66 2.3 ...
```

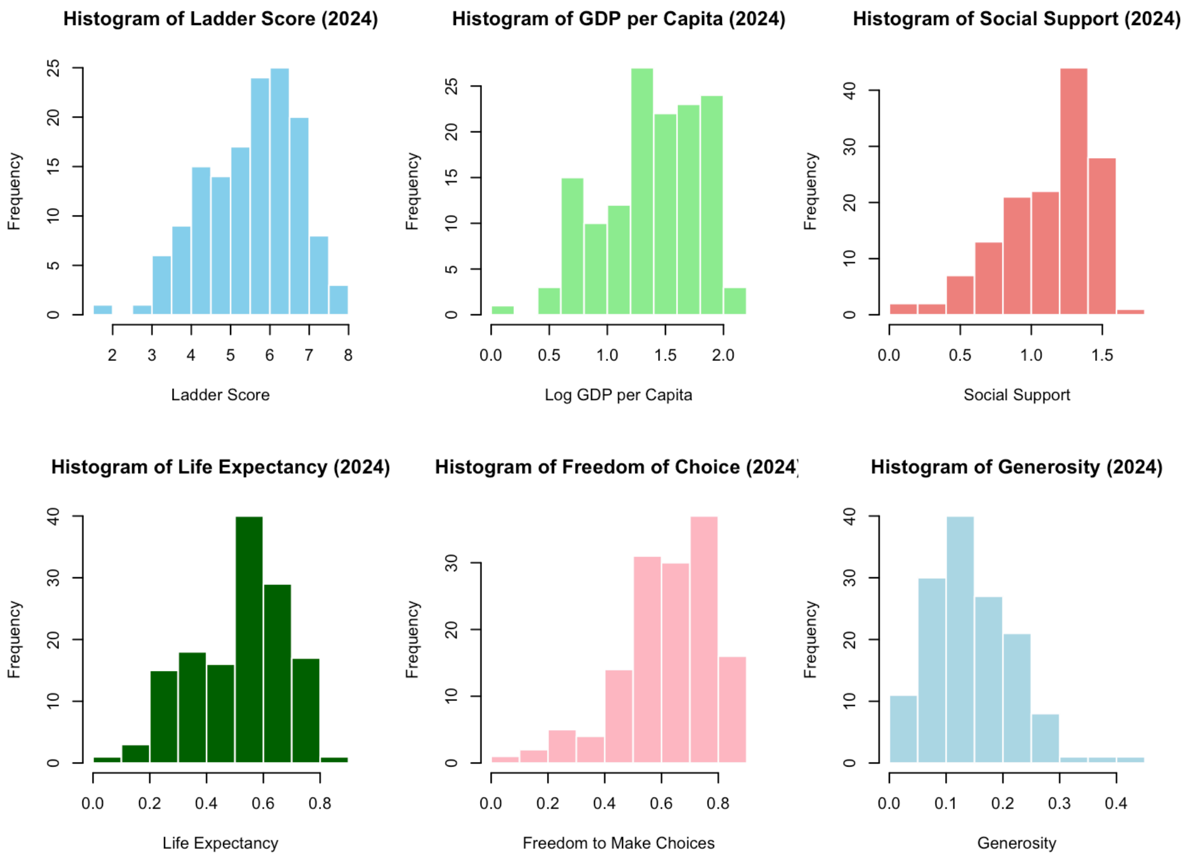
```
## 'data.frame':    2363 obs. of  11 variables:
##  $ Country.name                    : chr  "Afghanistan" "Afghanistan" "Afghanistan"
"Afghanistan" ...
##  $ year                            : int  2008 2009 2010 2011 2012 2013 2014 2015
2016 2017 ...
##  $ Life.Ladder                     : num  3.72 4.4 4.76 3.83 3.78 ...
##  $ Log.GDP.per.capita              : num  7.35 7.51 7.61 7.58 7.66 ...
##  $ Social.support                  : num  0.451 0.552 0.539 0.521 0.521 0.484 0.526
0.529 0.559 0.491 ...
##  $ Healthy.life.expectancy.at.birth: num  50.5 50.8 51.1 51.4 51.7 ...
##  $ Freedom.to.make.life.choices    : num  0.718 0.679 0.6 0.496 0.531 0.578 0.509
0.389 0.523 0.427 ...
##  $ Generosity                      : num  0.164 0.187 0.118 0.16 0.234 0.059 0.102
0.078 0.04 -0.123 ...
##  $ Perceptions.of.corruption       : num  0.882 0.85 0.707 0.731 0.776 0.823 0.871
0.881 0.793 0.954 ...
##  $ Positive.affect                 : num  0.414 0.481 0.517 0.48 0.614 0.547 0.492
0.491 0.501 0.435 ...
##  $ Negative.affect                 : num  0.258 0.237 0.275 0.267 0.268 0.273 0.375
0.339 0.348 0.371 ...
```

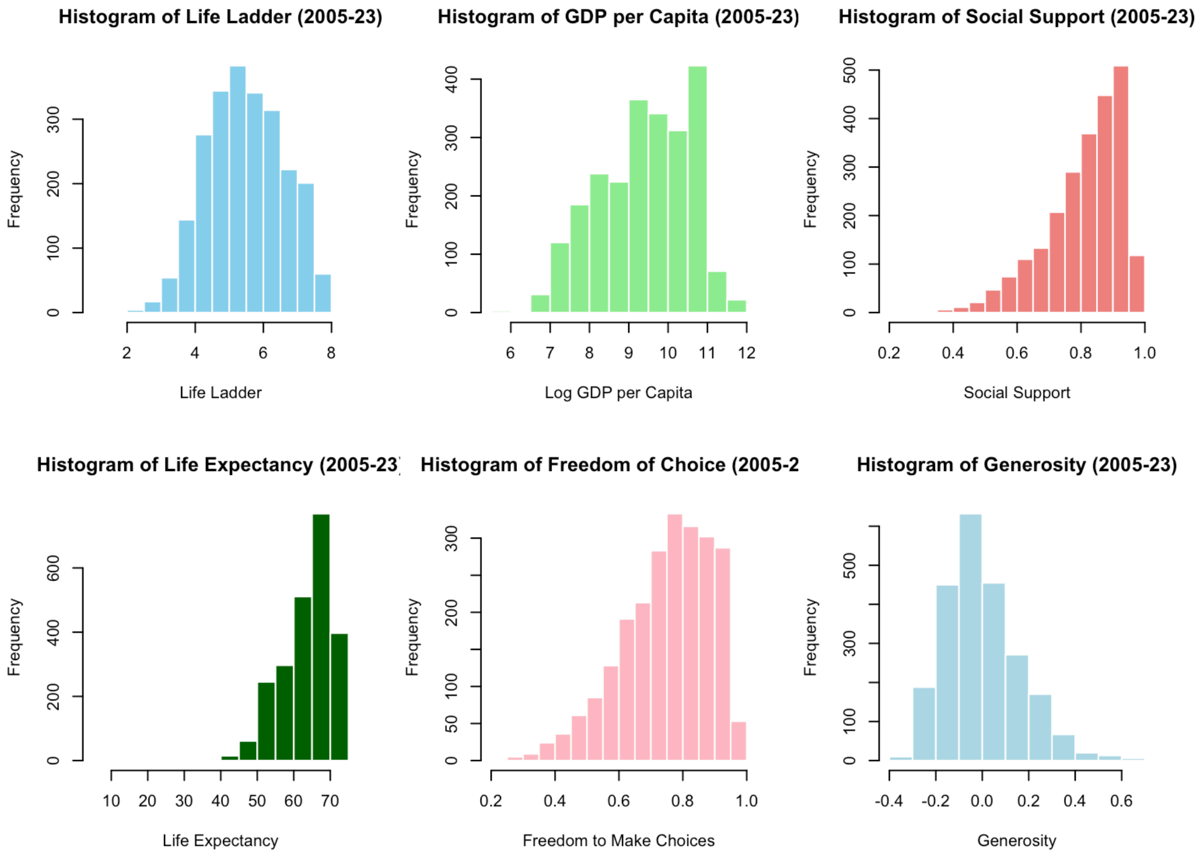
```
##   Country.name           Regional.indicator Ladder.score upperwhisker
## 1      Finland               Western Europe        7.741        7.815
## 2      Denmark               Western Europe        7.583        7.665
## 3      Iceland               Western Europe        7.525        7.618
## 4       Sweden               Western Europe        7.344        7.422
## 5       Israel Middle East and North Africa        7.341        7.405
## 6  Netherlands               Western Europe        7.319        7.383
##   lowerwhisker Log.GDP.per.capita Social.support Healthy.life.expectancy
## 1        7.667              1.844          1.572                   0.695
## 2        7.500              1.908          1.520                   0.699
## 3        7.433              1.881          1.617                   0.718
## 4        7.267              1.878          1.501                   0.724
## 5        7.277              1.803          1.513                   0.740
## 6        7.256              1.901          1.462                   0.706
##   Freedom.to.make.life.choices Generosity Perceptions.of.corruption
## 1                        0.859      0.142                     0.546
## 2                        0.823      0.204                     0.548
## 3                        0.819      0.258                     0.182
## 4                        0.838      0.221                     0.524
## 5                        0.641      0.153                     0.193
## 6                        0.725      0.247                     0.372
##   Dystopia...residual
## 1               2.082
## 2               1.881
## 3               2.050
## 4               1.658
## 5               2.298
## 6               1.906
```

```
##   Country.name year Life.Ladder Log.GDP.per.capita Social.support
## 1  Afghanistan 2008       3.724              7.350          0.451
## 2  Afghanistan 2009       4.402              7.509          0.552
## 3  Afghanistan 2010       4.758              7.614          0.539
## 4  Afghanistan 2011       3.832              7.581          0.521
## 5  Afghanistan 2012       3.783              7.661          0.521
## 6  Afghanistan 2013       3.572              7.680          0.484
##   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices Generosity
## 1                             50.5                        0.718      0.164
## 2                             50.8                        0.679      0.187
## 3                             51.1                        0.600      0.118
## 4                             51.4                        0.496      0.160
## 5                             51.7                        0.531      0.234
## 6                             52.0                        0.578      0.059
##   Perceptions.of.corruption Positive.affect Negative.affect
## 1                     0.882           0.414           0.258
## 2                     0.850           0.481           0.237
## 3                     0.707           0.517           0.275
## 4                     0.731           0.480           0.267
## 5                     0.776           0.614           0.268
## 6                     0.823           0.547           0.273
```

3.3. Visualize data distributions

Visualize the data with histogram to understand the distribution of key variables like Ladder score, GDP per capita, Social Support, etc., in both the datasets. This step provides a summary of the general trends in the data through histogram distribution plots.





Analysis: The histograms provide a visual summary of the key variables from the two datasets, reflecting the overall trends and distributions. From the histograms it can be ascertained that, the distribution of ladder score is roughly symmetric which implies that, very few countries have extremely high or extremely low ladder scores in both the datasets. However, the histograms of Log GDP per Capita of both the datasets are observed to be right-skewed (positively skewed) which highlights that most countries have been in the lower GDP per capita range, with very few countries being significantly wealthy.

In summary, the two datasets reveal trends in global well-being, with 2024 reflecting relatively high scores in happiness, social support, and life expectancy. However, over the longer period from 2005-2023, there’s more variability, especially in social support and life expectancy, suggesting that global trends have fluctuated over time, possibly due to varying global events and changes. This variability could be attributed to different global events and changes over the years, highlighting the importance of understanding these trends in the context of the data. However, the nature of non-normality seen in certain variables like, Log GDP per capita or Life expectancy will be further investigated in the next steps.

3.4. Data cleaning

This step identifies and addresses any missing values in the datasets. This is because missing values can distort the observed patterns and trends. Therefore, it's essential to identify and address missing data to maintain the integrity of this Exploratory Data Analysis (EDA).

```
##   Log.GDP.per.capita Social.support Healthy.life.expectancy
## 1                  3              3                       3
##   Freedom.to.make.life.choices Generosity Perceptions.of.corruption
## 1                            3          3                         3
##   Dystopia...residual
## 1                   3
```

```
##   Log.GDP.per.capita Social.support Healthy.life.expectancy.at.birth
## 1                 28             13                               63
##   Freedom.to.make.life.choices Generosity Perceptions.of.corruption
## 1                           36         81                       125
##   Positive.affect Negative.affect
## 1              24              16
```

3.5. Handle missing values

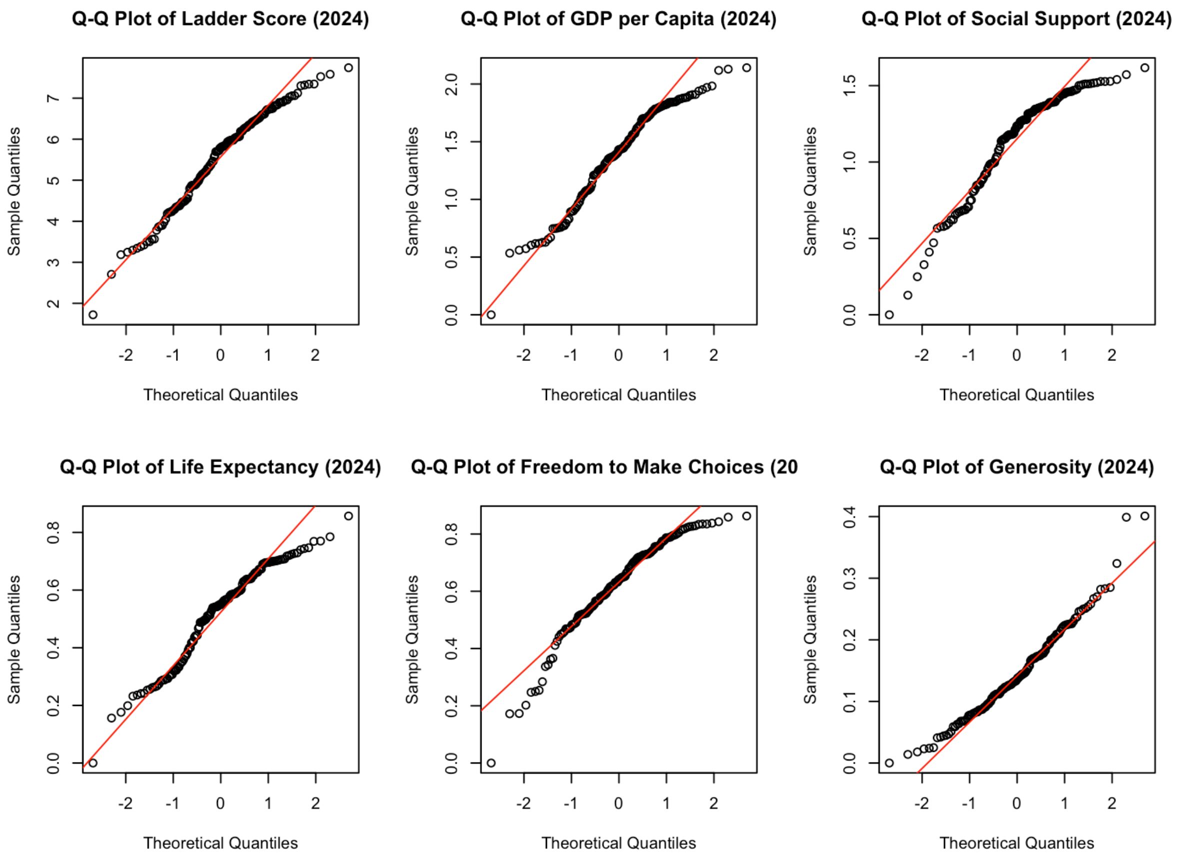
To handle missing values, 'NAs' and 'NaNs' are replaced with mean of the respective columns to ensure data integrity and consistency for further analysis. Here, 'grouping' ensures that calculating the 'mean' and filling in missing values are done within the context of each country, maintaining the integrity of country-specific data. And 'ungrouping' allows the 'mean' to be applied across the entire dataset, which is necessary for filling in any gaps left by the grouped operation.

```
## # A tibble: 6 × 12
##   Country.name Regional.indicator        Ladder.score upperwhisker lowerwhisker
##   <chr>        <chr>                            <dbl>        <dbl>        <dbl>
## 1 Finland      Western Europe                    7.74         7.82         7.67
## 2 Denmark      Western Europe                    7.58         7.66         7.5 
## 3 Iceland      Western Europe                    7.52         7.62         7.43
## 4 Sweden       Western Europe                    7.34         7.42         7.27
## 5 Israel       Middle East and North Afr…        7.34         7.40         7.28
## 6 Netherlands  Western Europe                    7.32         7.38         7.26
## # ℹ 7 more variables: Log.GDP.per.capita <dbl>, Social.support <dbl>,
## #   Healthy.life.expectancy <dbl>, Freedom.to.make.life.choices <dbl>,
## #   Generosity <dbl>, Perceptions.of.corruption <dbl>,
## #   Dystopia...residual <dbl>
```

```
##   Country.name year Life.Ladder Log.GDP.per.capita Social.support
## 1  Afghanistan 2008       3.724              7.350          0.451
## 2  Afghanistan 2009       4.402              7.509          0.552
## 3  Afghanistan 2010       4.758              7.614          0.539
## 4  Afghanistan 2011       3.832              7.581          0.521
## 5  Afghanistan 2012       3.783              7.661          0.521
## 6  Afghanistan 2013       3.572              7.680          0.484
##   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices Generosity
## 1                             50.5                        0.718      0.164
## 2                             50.8                        0.679      0.187
## 3                             51.1                        0.600      0.118
## 4                             51.4                        0.496      0.160
## 5                             51.7                        0.531      0.234
## 6                             52.0                        0.578      0.059
##   Perceptions.of.corruption Positive.affect Negative.affect
## 1                     0.882           0.414           0.258
## 2                     0.850           0.481           0.237
## 3                     0.707           0.517           0.275
## 4                     0.731           0.480           0.267
## 5                     0.776           0.614           0.268
## 6                     0.823           0.547           0.273
```

3.6. Normality Check Using Q-Q Plot

The histogram visualization performed in Step 3 gave a quick sense of how data is spread across different values and showed how certain varibles deviated from the normal distribution. However, to specifically understand where the deviations occurs, a Normality check using Q-Q plot is performed in this step. These Q-Q plots serve as a follow-up to confirm the normality or non-normality suggested by the histograms. This step is a validation for the analysis made in Step 3.



Q-Q Plot of Ladder Score (2024)

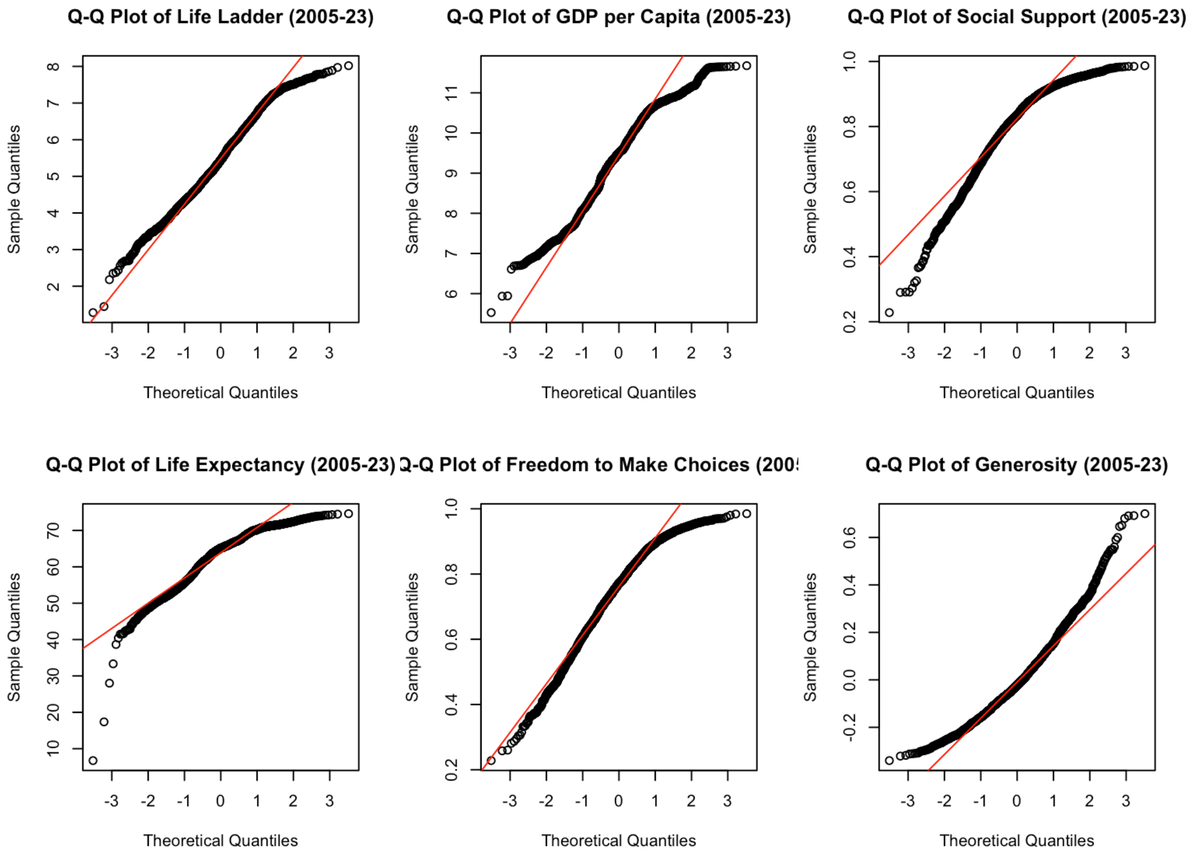
Q-Q Plot of GDP per Capita (2024)

Q-Q Plot of Social Support (2024)

Q-Q Plot of Life Expectancy (2024)

Q-Q Plot of Freedom to Make Choices (20

Q-Q Plot of Generosity (2024)



Q-Q Plot of Life Ladder (2005-23)

Q-Q Plot of GDP per Capita (2005-23)

Q-Q Plot of Social Support (2005-23)

Q-Q Plot of Life Expectancy (2005-23)

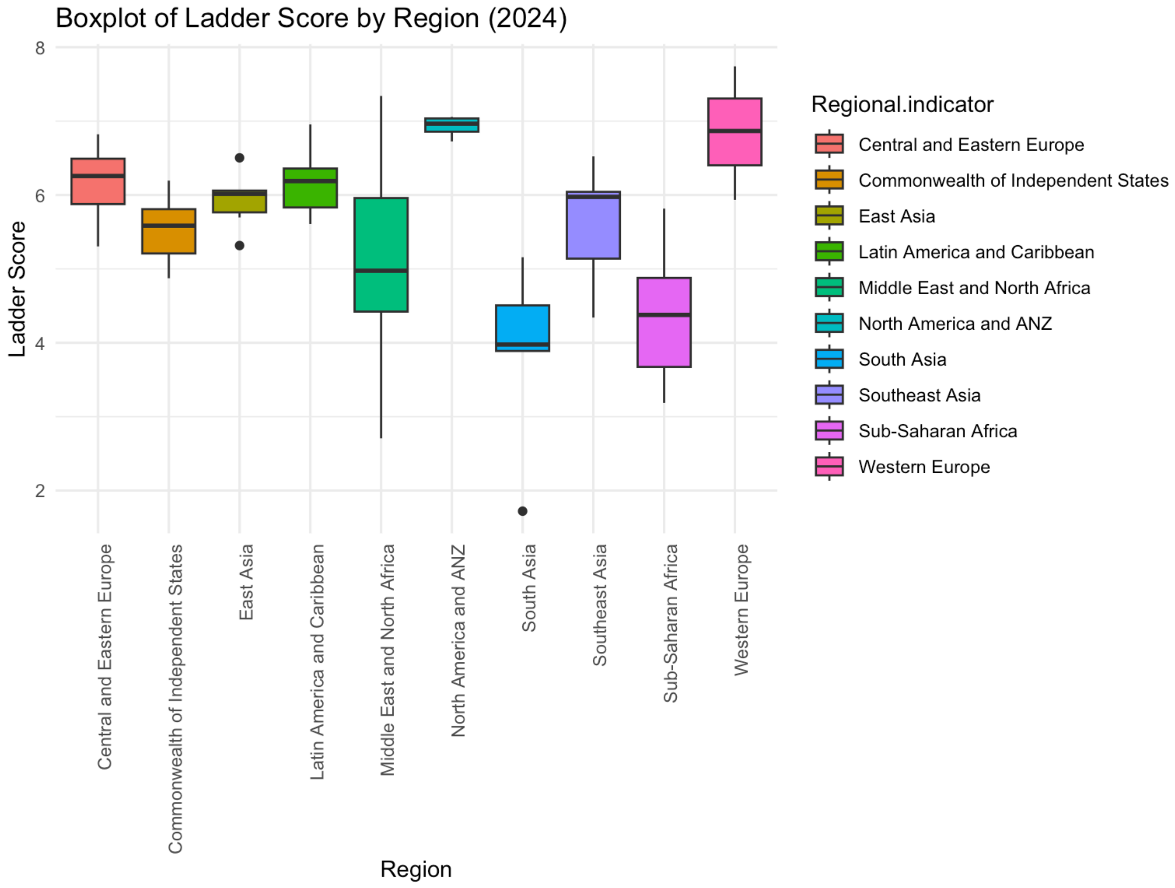
Q-Q Plot of Freedom to Make Choices (200

Q-Q Plot of Generosity (2005-23)

Analysis: The Q-Q plots show that Ladder Scores in both the datasets mostly follows the red line (theoretical quantile) but there are slight deviations at both the ends (tails). This suggests a mostly normal distribution. Further, for GDP per capita, the Q-Q plots of both the datasets show some deviation from the line in the lower tail, which indicates the presence of lower-than-expected GDP values that don't align with the normal distribution. In summary, the Q-Q plots confirmed the findings from the histograms, providing a detailed visualization of how some variables in the data closely align with a normal distribution while few others seem to have a slight deviation from the normal (mostly right skewed).

3.7. Box Plot Analysis

In the previous step, the Q-Q plots provided an overall assessment of the distribution of Ladder Score and other variables in the datasets. However, they did not offer insights into how these distributions differ across specific regions. Therefore, Box plots are employed in this step to specifically visualize the distribution of the Ladder Score across different regions in the 2024 dataset. This transition from a general examination of distributional normality to a more detailed investigation allows for the identification of regional variations and any outliers within the distribution.

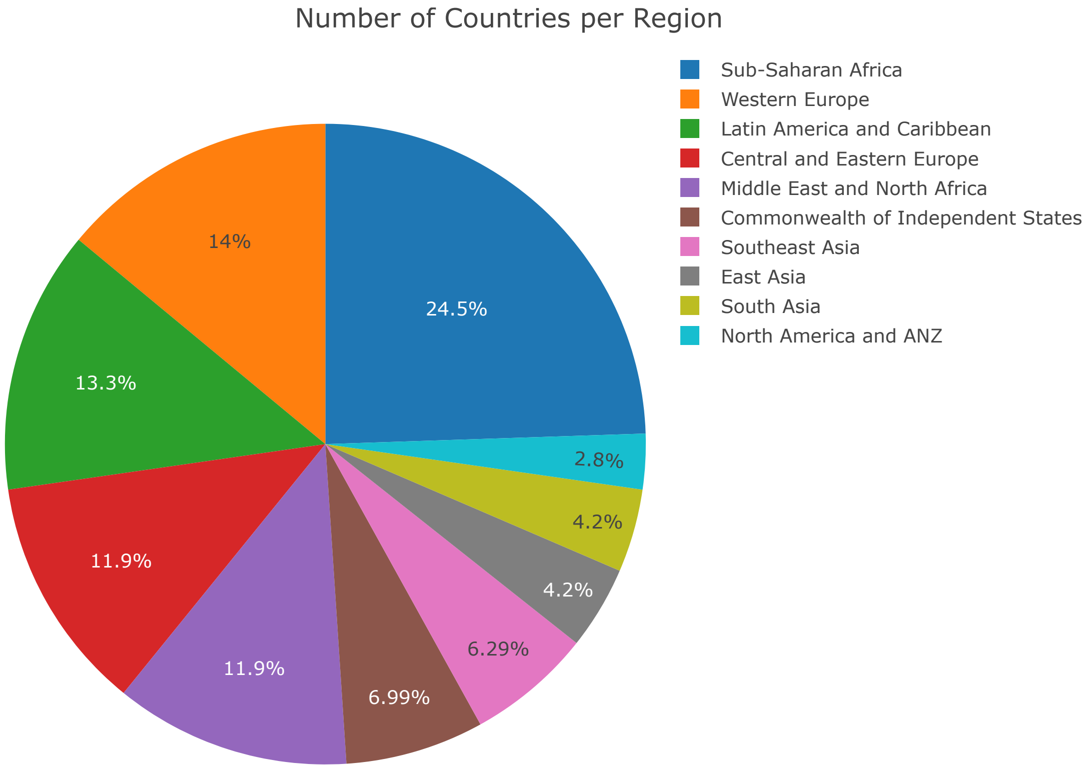


Analysis: From the above boxplot visualization, we can identify that Western Europe has the highest ladder score and Sub-Saharan Africa has the lowest ladder score. Further, we can also identify that Middle East and North Africa has the largest spread in the distribution of the ladder score. On the contrary, North America and ANZ has the lowest spread. However, it is essential to understand the underlying structure of the data in terms of how many countries are contributing to each region's score.

3.8. Distribution of Countries by Region

The Box plots in the previous step, gave insights into the spread and central tendency of the Ladder Scores across regions, but they did not reveal whether these observations were based on a large or small number of countries, which could potentially affect the reliability and generalizability of the findings.

By generating a Pie Chart that shows the distribution of countries by region, it is easier to visualize the number of countries represented in each region. This step adds context to the Box plot analysis by revealing the size of the sample from each region and providing an understanding of whether the observed trends in the Box plots are based on a large or small number of data points.

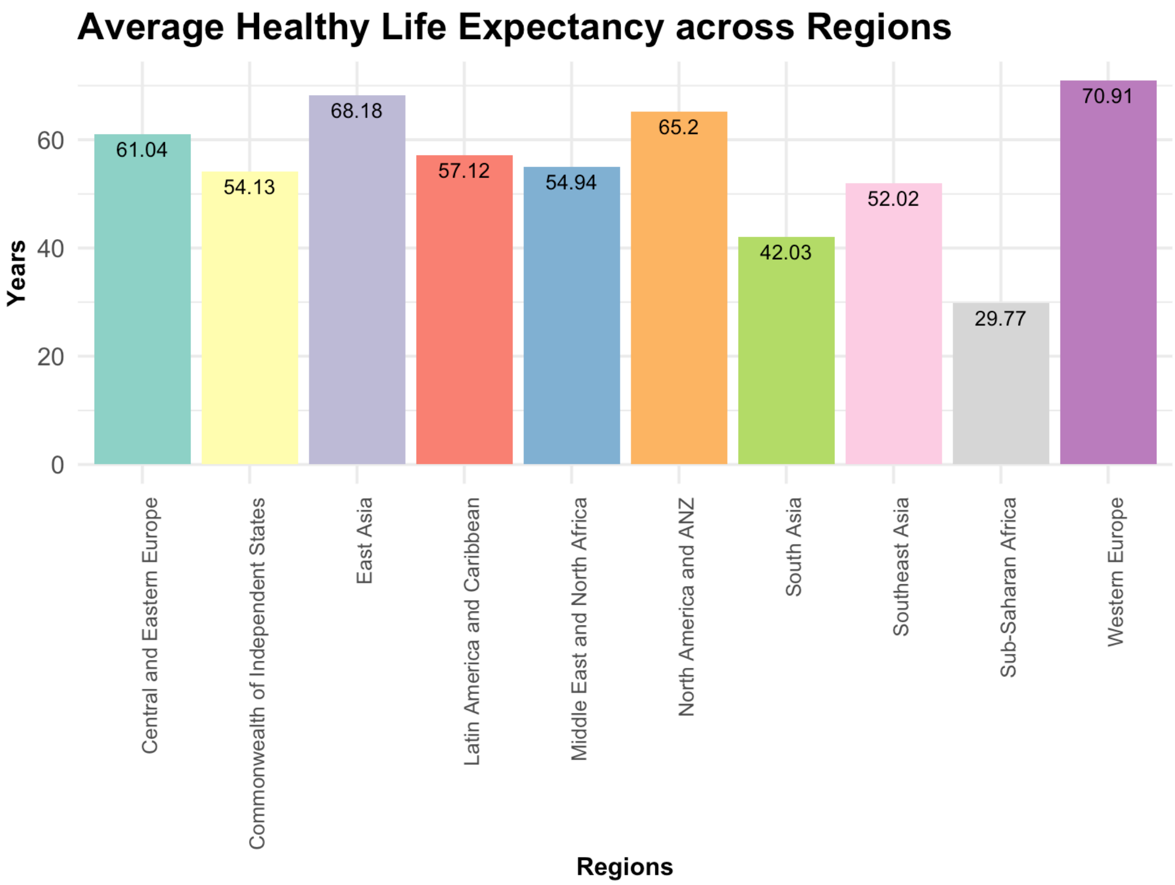


Analysis: The pie chart reveals that Sub-Saharan Africa constitutes the largest proportion of countries among all regions, accounting for approximately 24.5% of the total sample. When this is correlated with the box plot findings, which showed that Sub-Saharan Africa has the lowest Ladder Scores, it suggests that a significant portion of the global dataset is represented by countries in this region, many of which report lower happiness scores. Likewise, Western Europe is represented by a considerable portion of the countries (14%) in the dataset, as shown in the pie chart. This substantial representation, along with the high Ladder Scores observed in the box plot for Western Europe in the previous step, suggests that the positive trends in happiness in this region are not isolated to a few countries but are instead reflective of a broader regional trend. However, the pie chart reveals that North America and ANZ have the lowest representation in the dataset, making up only 2.8% of the total countries. This region displayed a low spread in Ladder Scores in the previous step, which indicates a relatively consistent level of happiness within the region. However, the small sample size, as shown in the Pie Chart, suggests that these findings may require to be interpreted with caution.

3.9. Mean Healthy Life Expectancy Across Regions

In the previous steps, the analysis of the distribution of Ladder Scores and the number of countries per region provided insights into the overall happiness and its variability across different regions. However, happiness could be influenced by various factors, including health. So, to gain a deeper understanding of the regional differences in well-being, it is essential to examine 'Healthy Life Expectancy at birth', which is a critical component of overall quality of life and well-being. Healthy life expectancy is calculated as the average number of years a newborn infant would live in good health. The plot of 'Mean Healthy Life Expectancy across regions' adds another layer of understanding by highlighting the disparities in health, which could explain some of the observed differences in the Ladder Scores across different regions.

This step will establish a correlation between happiness and health, and provides a more holistic view of regional well-being. The analysis also helps reveal whether regions with higher Ladder Scores also have longer healthy life expectancy, providing further evidence for the relationship between health and happiness.



Analysis: The bar plot shows that Western Europe has the highest average Healthy Life Expectancy, which aligns with its high Ladder Scores observed in the box plots. This suggests that the high happiness levels in Western Europe could be partly attributed to better health and longer life expectancy. Likewise, North America and ANZ also have a high average Healthy Life Expectancy, which corresponds to the relatively high Ladder Scores in this region. This supports the notion that regions with higher life expectancies tend to have higher happiness levels. However, Sub-Saharan Africa has the lowest Healthy Life Expectancy, which is consistent with its low Ladder Scores observed in the box plots. The combination of poor health and low life expectancy could be significant factors contributing to the lower Ladder Scores/ happiness levels in this region.

In summary, the analysis of Healthy Life Expectancy provides crucial context for understanding the regional disparities in ladder scores (happiness) observed in the earlier steps. Regions with higher life expectancy generally exhibit higher ladder scores or happiness levels, reinforcing the importance of health as a key component of well-being.

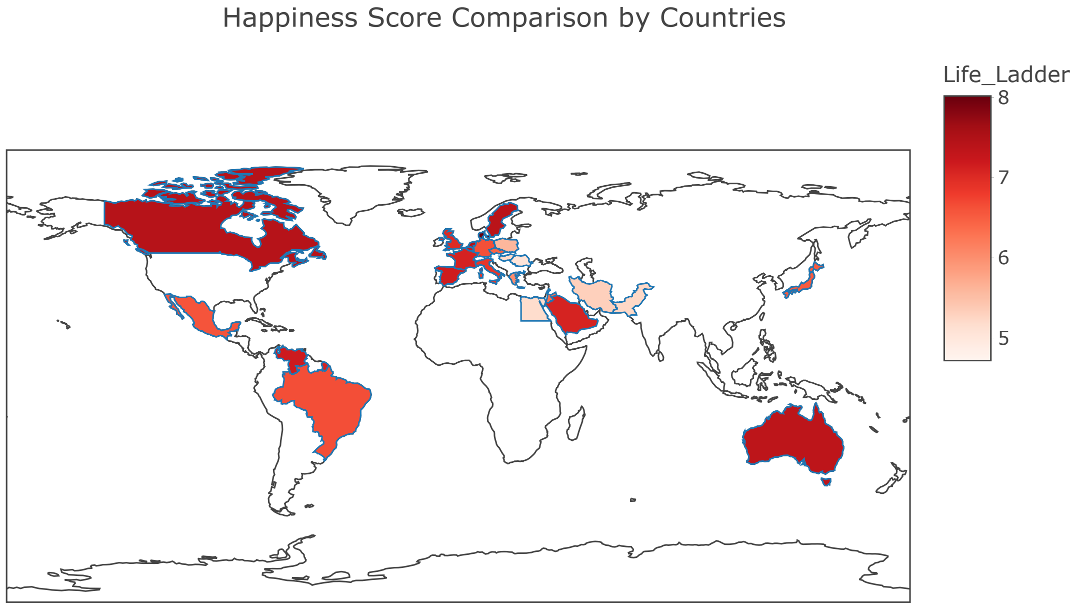
3.10. Analyze global happiness scores over entire time period

In the previous steps, the analysis focused on regional trends and correlations between happiness, health, and other socio-economic factors. These analyses were primarily region-based, offering a high-level view of the data. The interactive choropleth map in this step provides a dynamic perspective on how happiness has changed over time across different countries, thereby offering insights into the overall trends in global happiness. This map offers a different perspective by emphasizing geographic distribution and temporal changes. By introducing the dimension of time (in years), the choropleth map enables an analysis of how happiness scores have evolved over time within specific countries.

Pre-processing: When the analysis moves to a country-level comparison, particularly when visualizing data on a map (such as a choropleth map), the accuracy and consistency of country names become crucial. So, this data cleaning step is performed to ensure that the 'Country.name' column in the dataset is free of any encoding issues, non-ASCII characters, and other problematic strings that could cause errors during analysis or visualization. By setting the locale, converting text encoding, handling missing values, and removing invalid characters, the dataset is prepared for accurate and error-free plotting.

```
## [1] "en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8"
```

```
## [1] 0
```



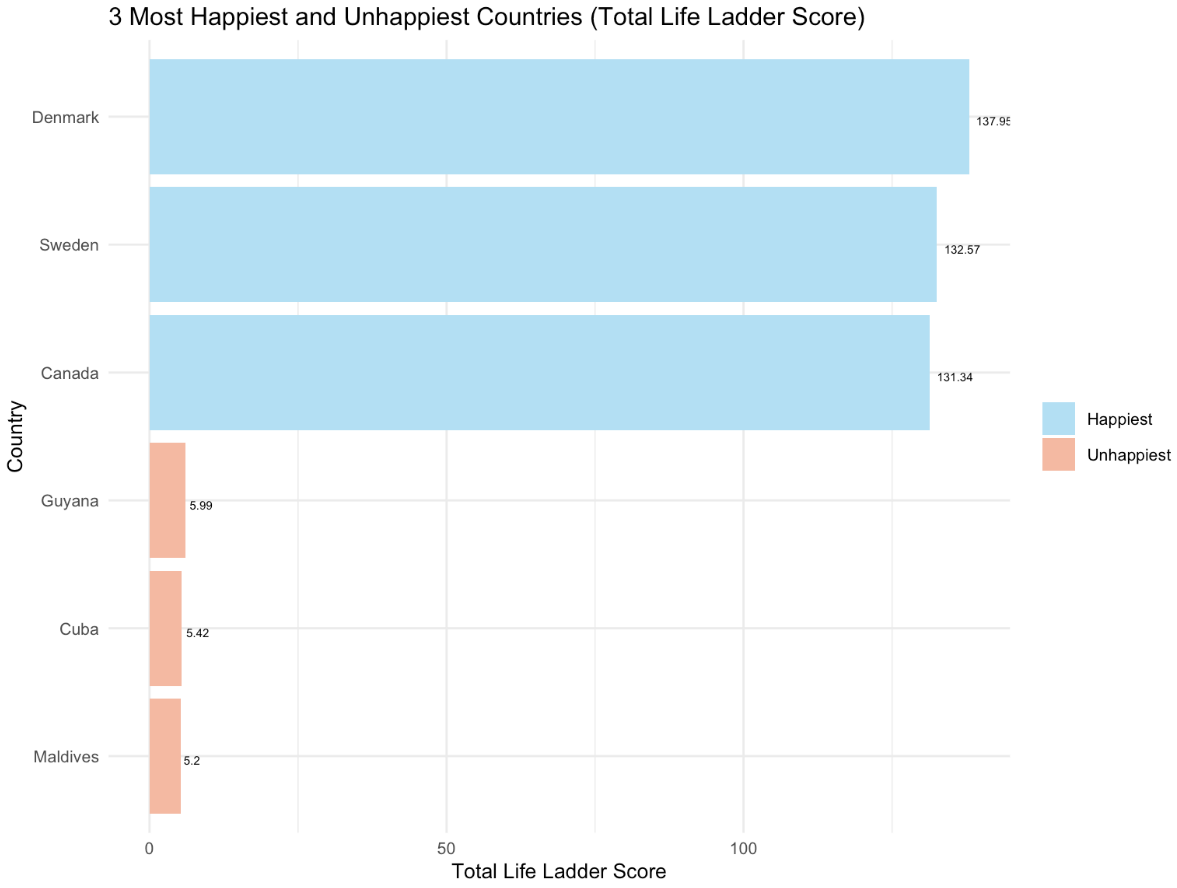
Happiness Score Comparison by Countries

Analysis: This interactive choropleth map provides the flexibility to observe the trends in Ladder scores across time (2005 - 2023) over different countries. In 2005, Denmark had the highest happiness score, followed by other happy countries such as Canada, Australia, Saudi Arabia, Spain, France, Sweden, the UK, and Venezuela. Conversely, Romania had the lowest happiness score, with Hungary, Iran, Pakistan, Poland, and Egypt also exhibiting low scores. In 2006, although most of the countries observed in the previous year are not visible, Venezuela still maintains a relatively high happiness score, albeit slightly lower than the previous year. Togo and Benin in West Africa have some of the lowest happiness scores, with 3.20 and 3.33, respectively. From 2007 to 2010, Canada consistently displayed high happiness scores, making it one of the happiest countries. Denmark, despite experiencing fluctuations, continued to hold the top spot until 2010 with a Ladder Score of 7.7.

However, by 2015, Denmark dropped to the third position on the list of happiest countries, with Norway and Switzerland taking the first and second positions, respectively. Liberia and Yemen emerged as the least happy countries that year. In 2021, Finland topped the list with a happiness score of 7.79, followed by Denmark and Iceland, while Afghanistan remained at the bottom with a score of 2.43. In 2022, Afghanistan's Ladder Score further decreased to 1.28, keeping it at the bottom of the list, while Finland maintained its position at the top with a score of 7.72. Although Afghanistan's Ladder Score improved slightly to 1.46 in 2023 compared to 2022, it still ranked lowest on the happiness scale. Finland, for the third consecutive year since 2021, remained the happiest country.

3.11. Comparative Analysis of Happiness Scores - Identify Happiest and Unhappiest Countries

After the choropleth map, which provided a temporal and geographical overview of the Ladder scores across countries over time, the comparative analysis of happiness scores was performed to pinpoint specific countries that consistently rank as the happiest and unhappiest based on their cumulative Ladder scores. It transitions from broad, dynamic visualizations to a more focused analysis, allowing for the identification of specific outliers (both positive and negative) in terms of happiness.



Analysis of Bar Plots: The bar plot reveals the top three happiest and unhappiest countries based on their cumulative Ladder scores. Denmark, Sweden, and Canada have the highest cumulative happiness scores, indicating that these countries have maintained consistently high happiness levels over the years. While Guyana, Cuba, and Maldives have the lowest cumulative happiness scores, suggesting that these countries have consistently low happiness levels across the dataset's timespan. These findings differ from year-by-year analyses shown in the choropleth map, where countries might temporarily rise or fall in happiness rankings due to short-term factors that might not affect their long-term cumulative scores.

In summary, these bar plots show a wide contrast in the happiness scores between the top 3 happiest and unhappiest countries, and this sets the stage for further analysis into the factors driving these differences.

3.12. Check for Normality

After identifying the happiest and unhappiest countries, it is essential to check the normality of the entire dataset and the subset of happiest and unhappiest countries to make appropriate selection of the statistical test to be performed. Understanding whether the data is normally distributed will guide the choice between parametric and non-parametric tests. And a Shapiro-Wilk normality test on the entire dataset ensures that the subsequent analysis is accurate and reliable. If the data is normally distributed, parametric tests such as the T-test, which have higher statistical power, can be used. If the data is not normally distributed, non-parametric tests would be more appropriate.

```
## 
##  Shapiro-Wilk normality test
## 
## data:  ds_2023_clean$Life.Ladder
## W = 0.98964, p-value = 4.935e-12
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  top_countries$Total_Life_Ladder
## W = 0.70486, p-value = 0.006953
```

Analysis: Since p-value is way less than 0.05 in the entire dataset and the combined dataset of the subsets - happiest and unhappiest countries, these datasets do not follow a normal distribution. In case of combined datasets, this could be because of merging two datasets that have very different central tendencies (means). The happiest countries have much higher Life Ladder scores compared to the unhappiest countries, which might lead to a skewed distribution when combined. This could in turn cause the Shapiro-Wilk test on the combined data to indicate non-normality.

However, focusing on the normally distributed subsets - the happiest and unhappiest countries individually, enables the advantages of parametric testing, such as greater power and precision, to be leveraged. This approach is crucial for understanding the sharp contrasts between the happiest and unhappiest countries. By isolating and analyzing these subsets, more meaningful comparisons between the extremes of the happiness spectrum can be made, providing clearer and more actionable insights.

```
## 
##  Shapiro-Wilk normality test
## 
```

```
## data:  top_3_happiest$Total_Life_Ladder
## W = 0.88385, p-value = 0.3358
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  top_3_unhappiest$Total_Life_Ladder
## W = 0.93768, p-value = 0.5181
```

Analysis of Shapiro-Wilk Normality test:The p-value is greater than 0.05 for Happiest and Unhappiest countries (p-value = 0.3358 and p-value = 0.5181), indicating that the data for both the happiest and unhappiest countries individually are normally distributed.

Thus, the data of Happiest and Unhappiest countries can be tested individually using a parametric test (T-test) since these two data when considered individually are normally distributed. T-test can be used to compare the mean happiness scores (Total Life Ladder) between the happiest and unhappiest countries.

3.13. Perform T- Test

In order to statistically validate the differences between the happiest and unhappiest countries, a T-test is performed.

```
##
##  One Sample t-test
##
## data:  top_3_happiest$Total_Life_Ladder
## t = 65.99, df = 2, p-value = 0.0002296
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  125.2224 142.6909
## sample estimates:
## mean of x
##  133.9567
```

```
##
##  One Sample t-test
##
## data:  top_3_unhappiest$Total_Life_Ladder
## t = 23.36, df = 2, p-value = 0.001828
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.516601 6.556066
## sample estimates:
## mean of x
##  5.536333
```

Analysis of the test: The Happiest Countries have a high mean happiness score (Total Life Ladder) of 133.96, which is statistically significant, as indicated by the low p-value (0.0002296) and a high t-value (65.99). This suggests that these countries consistently score very high on the happiness scale. However, the Unhappiest Countries have a much lower mean happiness score of 5.54, also statistically significant (p-value: 0.001828, t-value: 23.36). This confirms that these countries are at the opposite end of the happiness spectrum, but still have a score that is statistically above zero, indicating some level of happiness.

In summary, the T-test results show that the happiest countries have a significantly higher mean happiness score (Total Life Ladder) compared to the unhappiest countries. The T-test confirms that the differences between these two groups are not just due to random chance but are statistically significant.

3.14. Correlation Analysis of Factors Influencing Happiness

After confirming the significant difference in happiness scores between the happiest and unhappiest countries, the next step is to understand what drives these differences. This step analyzes the correlation between various factors (such as GDP per capita, social support, etc.) and the Life Ladder score for the happiest and unhappiest countries. Understanding these relationships will help uncover the key factors influencing happiness in these countries.

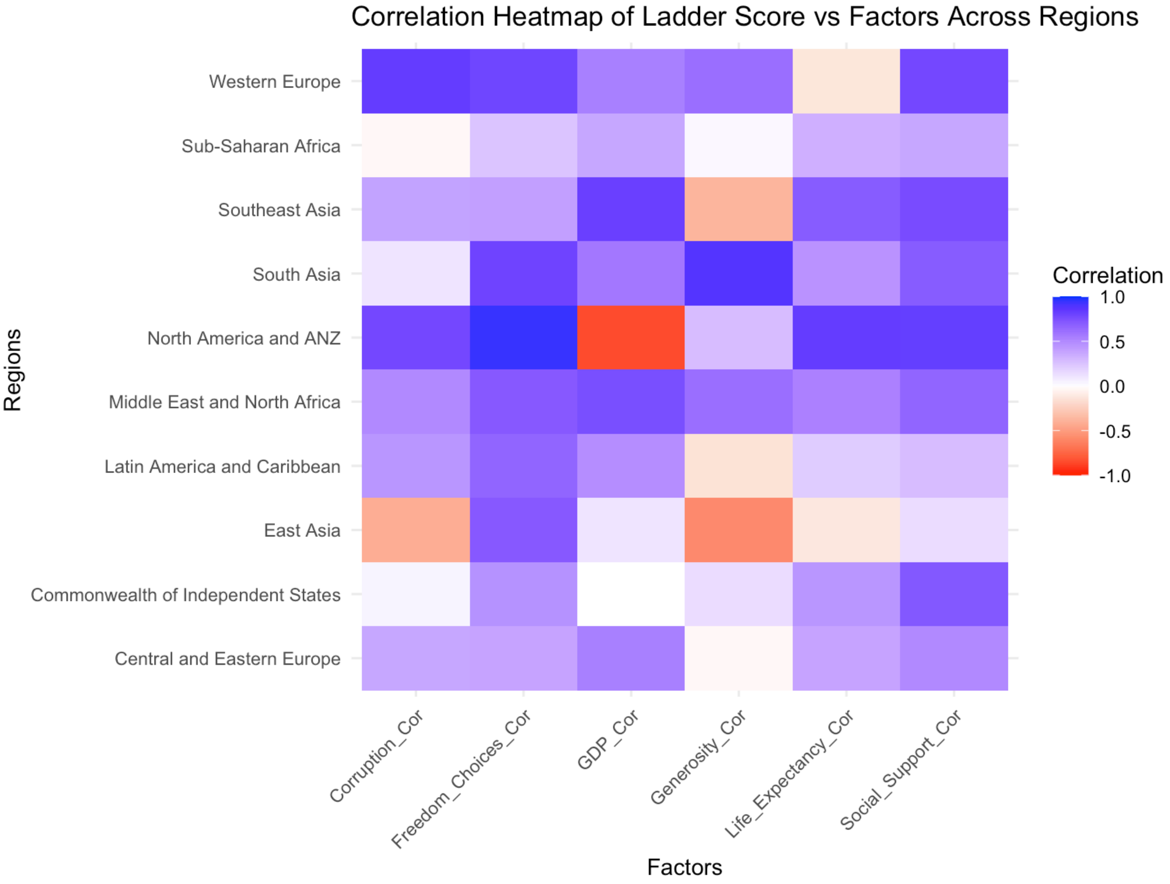
```
## [1] "Canada"   "Cuba"     "Denmark"  "Guyana"   "Maldives" "Sweden"
```

```
## # A tibble: 3 × 3
##   Country.name Factor                         Correlation
##   <chr>        <chr>                                <dbl>
## 1 Canada       Healthy.life.expectancy.at.birth      -0.836
## 2 Denmark      Healthy.life.expectancy.at.birth      -0.757
## 3 Sweden       Social.support                         0.354
```

Analysis: The analysis reveals that different factors significantly impact happiness across these countries. In Canada and Denmark, the unexpected negative correlation between Healthy Life Expectancy and happiness suggests that longevity alone does not guarantee happiness; other factors related to the quality of life, healthcare systems, and social conditions might play a crucial role. In contrast, Sweden's positive correlation between Social Support and happiness underscores the importance of social connections and community support in enhancing well-being. These findings highlight the complex and multifaceted nature of happiness, demonstrating that it is influenced by a variety of factors that can differ significantly from one country to another. Understanding these relationships helps in identifying key areas for policy intervention and improvement to enhance overall well-being in different regions.

3.15. Regional Correlation Analysis

After examining the correlation factors in the happiest and unhappiest countries, it’s important to expand the analysis to a regional level. This step explores how different factors correlate with happiness scores across various regions, helping to identify region-specific trends and influences.



Analysis: The heatmap reveals that the relationship between happiness and various socio-economic factors varies significantly across regions. Western Europe, North America, and ANZ consistently show positive correlations across multiple factors, suggesting that in these regions, improvements in economic and social conditions lead to higher happiness. On the other hand, regions like East Asia and Sub-Saharan Africa show weaker or even negative correlations with some factors, indicating that the drivers of happiness may be more complex or less directly linked to these specific variables. This heatmap underscores the importance of regional context in understanding the factors that influence happiness and suggests that a one-size-fits-all approach may not be appropriate when assessing global happiness metrics. It highlights the need for region-specific strategies to improve happiness and well-being.

3.16. Analysis of happiness score vs socio-economic factors

The heatmap in the previous step underscores the importance of regional context in understanding the factors that influence happiness and suggests that a one-size-fits-all approach may not be appropriate when assessing global happiness metrics. It highlights the need for region-specific strategies to improve happiness and well-being. To further explore these regional trends and the relationships between happiness and various socio-economic factors, scatter plots with regression lines are created. This step provides a more detailed visualization, allowing for the examination of linearity and the strength of these relationships, thereby confirming and complementing the findings from the correlation analysis.

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

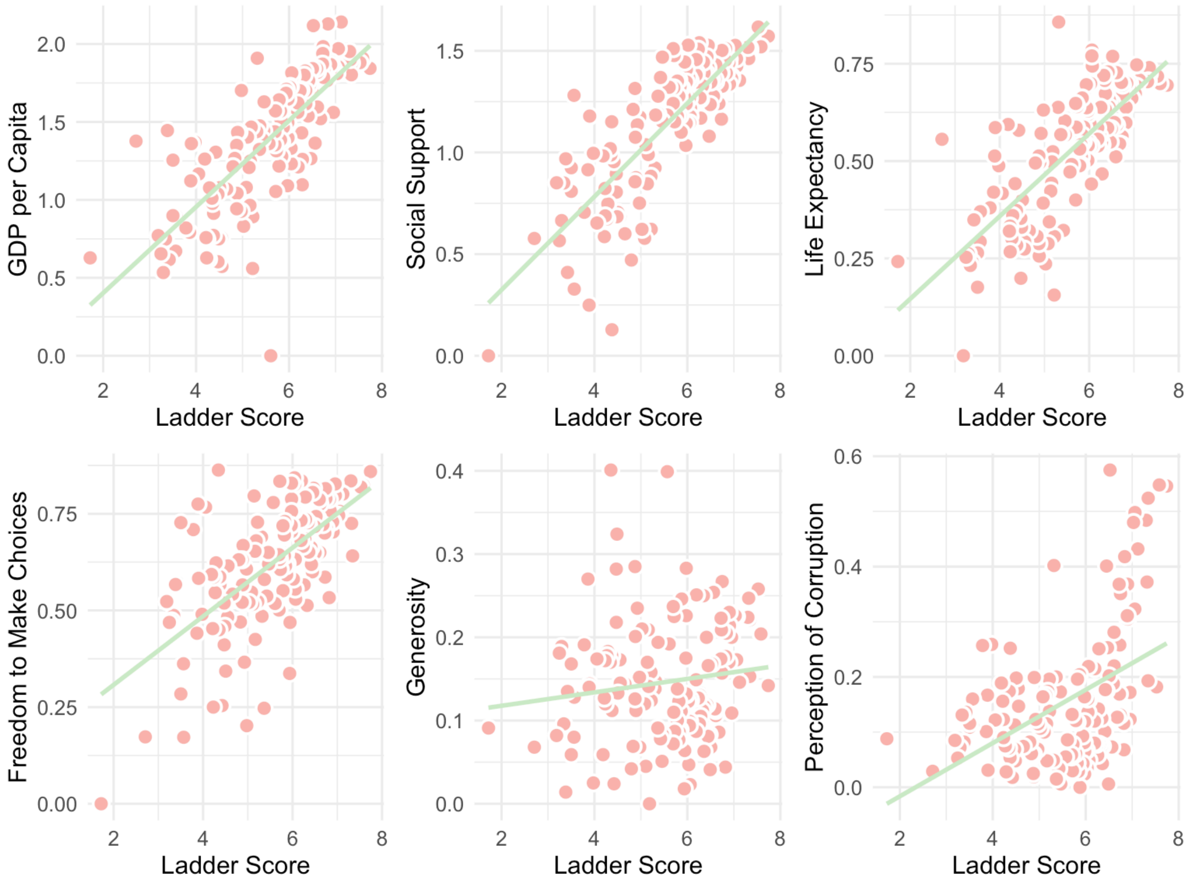
```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



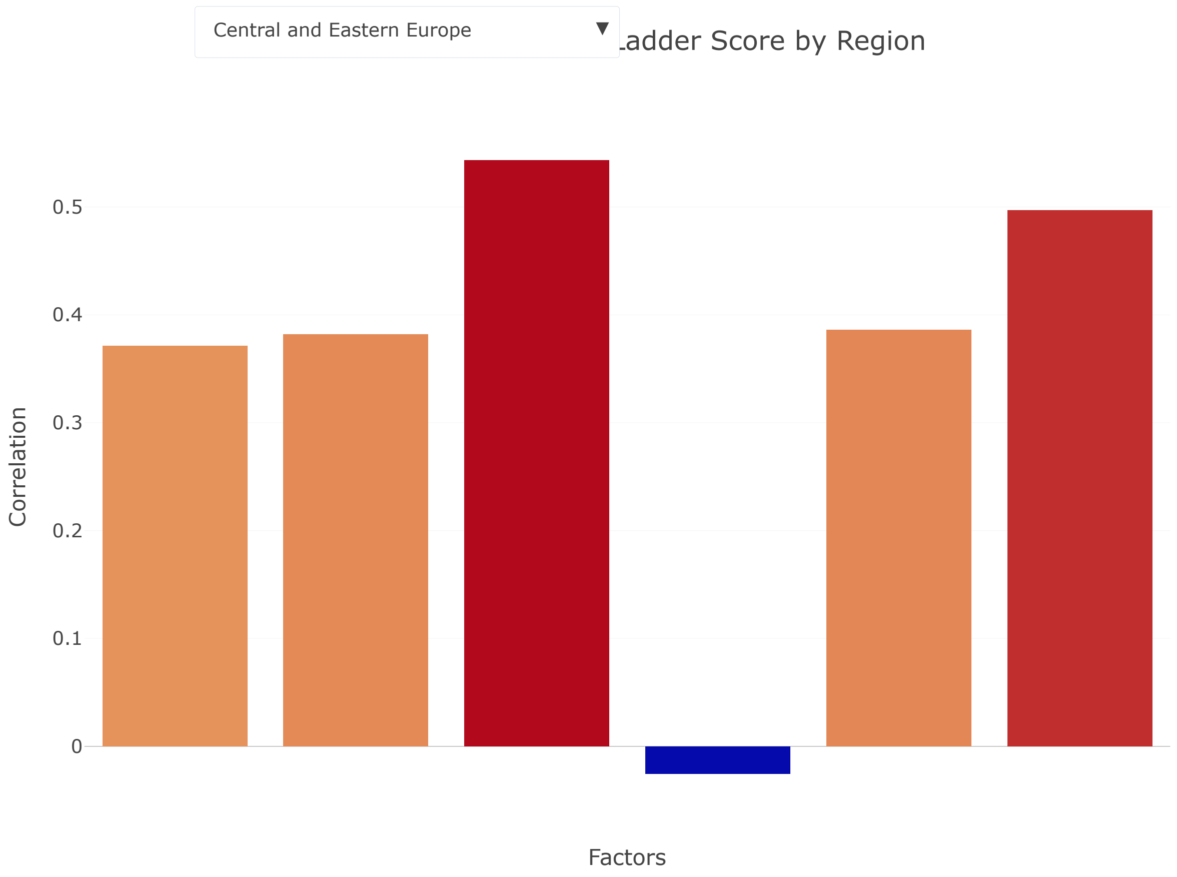
Analysis: The scatter plots with regression lines reveal the following relationships between happiness (Ladder score) and socio-economic factors:

GDP per Capita: A strong positive relationship, indicating that higher economic output per person is linked to greater happiness. Social Support: A positive correlation, suggesting that stronger social networks contribute significantly to higher happiness levels. Life Expectancy: A positive relationship, emphasizing that longer, healthier lives are associated with increased happiness. Freedom to Make Choices: A positive correlation, showing that greater autonomy and freedom in life choices lead to higher happiness. Generosity: A weaker correlation, indicating that while generosity contributes to happiness, it is less impactful than other factors like economic stability or social support. Perception of Corruption: A more complex and weaker positive relationship, suggesting that the impact of corruption on happiness varies and may be influenced by other mitigating factors.

3.17. Interactive dashboard to investigate global impact of socio-economic factors on happiness

To enhance the depth of analysis, an advanced interactive visualization is created. This interactive plot allows for a more granular exploration of how different factors influence happiness across regions. By enabling the examination of region-specific correlations, this visualization provides an opportunity to interact with the data dynamically, offering deeper insights into the complex relationships between happiness and its drivers.

Note: In slidey mode, every click, including those on dropdown menus, may advance to the next slide. After selecting variables from the dropdown, please manually navigate back to the page to explore the interactive dashboard.



Analysis: This interactive visualization provides a dynamic tool for exploring how different factors correlate with happiness across various regions. By allowing users to filter and compare regions, it offers a more detailed understanding of how regional contexts influence the relationship between different socio-economic factors and happiness.

Section 4: Conclusion and Future Scope

This exploratory data analysis (EDA) has provided a comprehensive examination of the factors influencing global happiness.Through a combination of statistical tests and visualizations, key insights into the drivers of happiness have been uncovered, highlighting the importance of regional context and the significant role of economic and social factors.The analysis also reveals the complexity of the relationships that contribute to overall well-being.

The advanced interactive visualization developed in this project serves as a powerful tool for further exploration, offering a dynamic way to delve deeper into the data. However, while the current analysis provides valuable insights, it also opens avenues for further exploration and refinement. Incorporating more sophisticated statistical techniques, such as multivariate regression or machine learning, could yield even deeper insights into the complex relationships between socio-economic factors and happiness.

Additionally, the interactive dashboard could be enhanced with features like multi-factor analysis, real-time updates, and predictive modeling to offer more dynamic and actionable insights.

This thorough EDA not only lays the foundation for more targeted research but also supports interventions aimed at improving happiness on both a global and regional scale.

References

1. Helliwell JF, Layard R, Sachs JD. World Happiness Report 2019.

2. Scatterplot with regression fit and automatic text repel – the R Graph Gallery [Internet]. [cited 2024 Aug 19]. Available from: https://r-graph-gallery.com/web-scatterplot-corruption-and-human-development.html

3. Happiness of the younger, the older, and those in between [Internet]. worldhappiness.report. Available from: https://worldhappiness.report/ed/2024/happiness-of-the-younger-the-older-and-those-in-between/

4. Tableau.com. 2024. Available from: https://public.tableau.com/app/profile/worldhappiness/viz/2024Draft/Figure2_

5. World Happiness Report. Explore the World Happiness Report data [Internet]. worldhappiness.report. 2023. Available from: https://worldhappiness.report/data/