
# PROGETTO DI STATISTICA E ANALISI DEI DATI

## Analisi del livello di felicità globale

**Autori**

Dashchuk Yulia (NF22500146)

Genovese Vincenzo (NF22500192)

Università degli studi di Salerno

a.a. 2025/2026

# Indice

# Introduzione

La misurazione della felicità e del benessere soggettivo rappresenta uno dei temi importanti per l'analisi della popolazione e del governo. Il Prodotto Interno Lordo (PIL), storicamente utilizzato come metro principale di valutazione dello sviluppo di un paese, si è rivelato insufficiente a descrivere la complessità delle condizioni di vita della popolazione, rendendo necessaria l'introduzione di fattori sociali e psicologici.

Il *World Happiness Report*[1] è un progetto di ricerca internazionale che misura la percezione della felicità nei diversi paesi del mondo e analizza i fattori socio-economici che possono influenzarla tramite sondaggi annuali su scala globale. La felicità viene quantificata attraverso una scala da 0 a 10 in cui gli individui valutano la soddisfazione della propria vita, permettendo così di analizzare un costrutto psicologico soggettivo, traducendolo in una misura statistica comparabile e integrandolo con indicatori oggettivi relativi alle condizioni economiche, sociali e politiche del paese.

L'analisi dei dati raccolti evidenzia come la felicità di un paese non sia un fenomeno monofattoriale ma il risultato dell'interazione fra i vari fattori soggettivi e oggettivi. Elementi come il supporto sociale, la percezione della corruzione e la libertà di compiere scelte importanti sono importanti per determinare il benessere di una popolazione. È essenziale comprendere quali di queste variabili contano di più, quanto influenzano il punteggio di felicità e come cambiano nel tempo.

È possibile applicare l'analisi della felicità mondiale in vari ambiti per evidenziare pattern geopolitici, differenze culturali, variazioni temporali e possibili effetti di crisi economiche o sanitarie (es. il COVID_19), per comprendere i fenomeni sociali e supportare decisioni politiche orientate al benessere. I governi potrebbero analizzare al meglio le aree su cui investire confrontando le scelte politiche di vari Paesi e notando i cambiamenti scatenati. Utilizzando i dati sulla felicità, inoltre, organizzazioni internazionali potrebbero programmare interventi mirati in nazioni con molti tratti negativi (criminalità, bassa libertà personale...).

[1] `https://www.worldhappiness.report`

# Analisi del dataset

Il dataset utilizzato [1] copre un arco temporale esteso (2005-2022) per ogni paese. Ognuna delle 2199 istanze rappresenta un paese in un dato anno e include 21 variabili (socio-economiche, demografiche ed emotive) sia in modo grezzo che standardizzato (Tabella 2.3). Le variabili, tranne il paese, il codice del paese e l'anno, sono quantitative e continue.
Nella tabella 2.1 sono riportate le prime dieci osservazioni del dataset, che mostrano i valori di felicità e i principali indicatori socio-economici per alcuni paesi e anni differenti.

| **Country** | **Code** | **Year** | **Happiness** | **Log GDP** | **Social** | **Health** | **Freedom** | **Generosity** | **Corruption** | **Pos. Affect** | **Neg. Affect** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | AFG | 2008 | 3.724 | 7.35 | 0.451 | 50.5 | 0.718 | 0.168 | 0.882 | 0.414 | 0.258 |
| Afghanistan | AFG | 2009 | 4.402 | 7.509 | 0.552 | 50.8 | 0.679 | 0.191 | 0.850 | 0.481 | 0.237 |
| Argentina | ARG | 2007 | 6.073 | 10.013 | 0.862 | 65.94 | 0.653 | -0.144 | 0.881 | 0.750 | 0.279 |
| Argentina | ARG | 2008 | 5.961 | 10.043 | 0.892 | 66.06 | 0.678 | -0.135 | 0.865 | 0.720 | 0.318 |
| Benin | BEN | 2015 | 3.625 | 7.955 | 0.434 | 54.3 | 0.733 | -0.026 | 0.850 | 0.555 | 0.373 |
| Benin | BEN | 2016 | 4.007 | 7.958 | 0.493 | 54.6 | 0.780 | -0.064 | 0.838 | 0.578 | 0.456 |
| Bulgaria | BGR | 2013 | 3.993 | 9.848 | 0.829 | 65.62 | 0.603 | -0.197 | 0.962 | 0.537 | 0.278 |
| Bulgaria | BGR | 2014 | 4.438 | 9.863 | 0.886 | 65.76 | 0.576 | -0.060 | 0.955 | 0.542 | 0.236 |
| Croatia | HRV | 2015 | 5.205 | 10.124 | 0.768 | 67.90 | 0.694 | -0.102 | 0.849 | 0.570 | 0.294 |
| Croatia | HRV | 2016 | 5.417 | 10.166 | 0.798 | 68.08 | 0.672 | -0.070 | 0.884 | 0.569 | 0.337 |

Tabella 2.1: Estratto del dataset: valori originali per i principali indicatori di felicità

Nel dataset sono inoltre riportati i valori standardizzati delle stesse variabili (Tabella 2.2), che ne rappresentano gli *z-score* ottenuti sottraendo la media e dividendo per la deviazione standard, in modo da consentire confronti diretti tra indicatori su scale differenti.
I valori standardizzati hanno media 0 e deviazione standard 1 e evidenziano quanto ciascun indicatore si discosti dalla media globale:

- Valori negativi indicano risultati inferiori alla media mondiale;
- Valori positivi rappresentano risultati superiori alla media;
- Le scale sono comparabili, permettendo di individuare pattern e correlazioni tra i diversi fattori.

[1] `https://raw.githubusercontent.com/m-clark/book-of-models/main/data/world_happiness_all_years.csv`

| Country | Year | Happiness_sc | GDP_sc | Social_sc | Health_sc | Freedom_sc | Generosity_sc | Corruption_sc | PosAffect_sc |
|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | 2008 | -1.56 | -1.77 | -2.97 | -1.85 | -0.21 | 1.04 | 0.74 | -2.25 |
| Afghanistan | 2009 | -0.96 | -1.63 | -2.14 | -1.81 | -0.49 | 1.19 | 0.56 | -1.62 |
| Argentina | 2007 | 0.53 | 0.54 | 0.42 | 0.38 | -0.68 | -0.89 | 0.73 | 0.92 |
| Argentina | 2008 | 0.43 | 0.57 | 0.67 | 0.40 | -0.50 | -0.84 | 0.64 | 0.64 |
| Benin | 2015 | -1.65 | -1.24 | -3.11 | -1.30 | -0.11 | -0.16 | 0.56 | -0.92 |
| Benin | 2016 | -1.31 | -1.24 | -2.63 | -1.26 | 0.23 | -0.40 | 0.50 | -0.70 |
| Bulgaria | 2013 | -1.32 | 0.40 | 0.15 | 0.34 | -1.03 | -1.22 | 1.17 | -1.09 |
| Bulgaria | 2014 | -0.93 | 0.41 | 0.62 | 0.36 | -1.23 | -0.37 | 1.13 | -1.04 |
| Croatia | 2015 | -0.24 | 0.64 | -0.35 | 0.67 | -0.38 | -0.63 | 0.56 | -0.78 |
| Croatia | 2016 | -0.06 | 0.67 | -0.10 | 0.69 | -0.54 | -0.44 | 0.75 | -0.79 |

Tabella 2.2: Estratto del dataset: Valori standardizzati

| N. | Nome attributo | Descrizione |
|---|---|---|
| 1 | `country` | Nome del Paese a cui si riferisce l'osservazione. |
| 2 | `cntry_code` | Codice ISO a tre lettere del Paese. |
| 3 | `year` | Anno di riferimento della rilevazione. |
| 4 | `happiness_score` | Punteggio medio di felicità percepita, su scala 0–10, derivato da sondaggi Gallup World Poll. |
| 5 | `log_gdp_per_capita` | PIL pro capite espresso in logaritmo naturale, come misura del benessere economico. |
| 6 | `social_support` | Valutazione del supporto sociale percepito: misura quanto le persone sentono di poter contare su qualcuno. |
| 7 | `healthy_life_expectancy_at_birth` | Aspettativa di vita in buona salute alla nascita. |
| 8 | `freedom_to_make_life_choices` | Indicatore che misura la libertà percepita di prendere decisioni di vita. |
| 9 | `generosity` | Indicatore della propensione alla generosità e alle donazioni verso gli altri. |
| 10 | `perceptions_of_corruption` | Percezione della corruzione nelle istituzioni pubbliche e private. Valori alti indicano maggiore corruzione percepita. |
| 11 | `positive_affect` | Frequenza di esperienze emotive positive. |
| 12 | `negative_affect` | Frequenza di emozioni negative. |
| 13 | `happiness_score_sc` | Punteggio di felicità standardizzato. |
| 14 | `log_gdp_per_capita_sc` | Valore standardizzato del PIL pro capite. |
| 15 | `social_support_sc` | Valore standardizzato del supporto sociale. |
| 16 | `healthy_life_expectancy_at_birth_sc` | Valore standardizzato dell'aspettativa di vita in buona salute. |
| 17 | `freedom_to_make_life_choices_sc` | Valore standardizzato della libertà di scelta nella vita. |
| 18 | `generosity_sc` | Indicatore standardizzato della generosità. |
| 19 | `perceptions_of_corruption_sc` | Valore standardizzato della percezione di corruzione. |
| 20 | `positive_affect_sc` | Valore standardizzato delle emozioni positive. |
| 21 | `negative_affect_sc` | Valore standardizzato delle emozioni negative. |

Tabella 2.3: Descrizione delle colonne del dataset

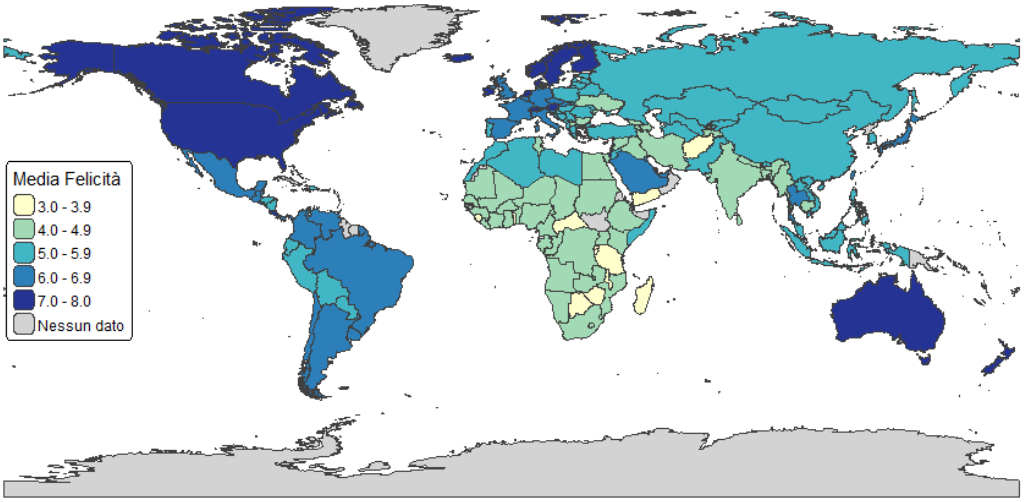


Figura 2.1: Mappa di felicità mondiale

## 2.1 Osservazioni sul dataset

### 2.1.1 Valori mancanti

| **Nome della variabile** | **Valori Mancanti** |
|---|---|
| country | 0 |
| cntry_code | 0 |
| year | 0 |
| happiness_score | 0 |
| log_gdp_per_capita | 20 |
| social_support | 13 |
| healthy_life_expectancy_at_birth | 54 |
| freedom_to_make_life_choices | 33 |
| generosity | 73 |
| perceptions_of_corruption | 116 |
| positive_affect | 24 |
| negative_affect | 16 |

Tabella 2.4: Numero di valori mancanti per ogni variabile originale del dataset

Dopo un'analisi del dataset si è osservata la presenza di valori mancanti distribuiti in modo non uniforme tra le variabili (tabella 2.4). In particolare le percentuali più alte di valori

mancanti si rilevano nell'Aspettativa di vita sana, Generosità e Percezione della corruzione. Ciò può essere dovuto principalmente a difficoltà di rilevazione in alcuni paesi (Figura 2.2). Risulta inoltre più difficile rilevare variabili soggettive rispetto a indicatori economici o demografici (Figura 2.3).

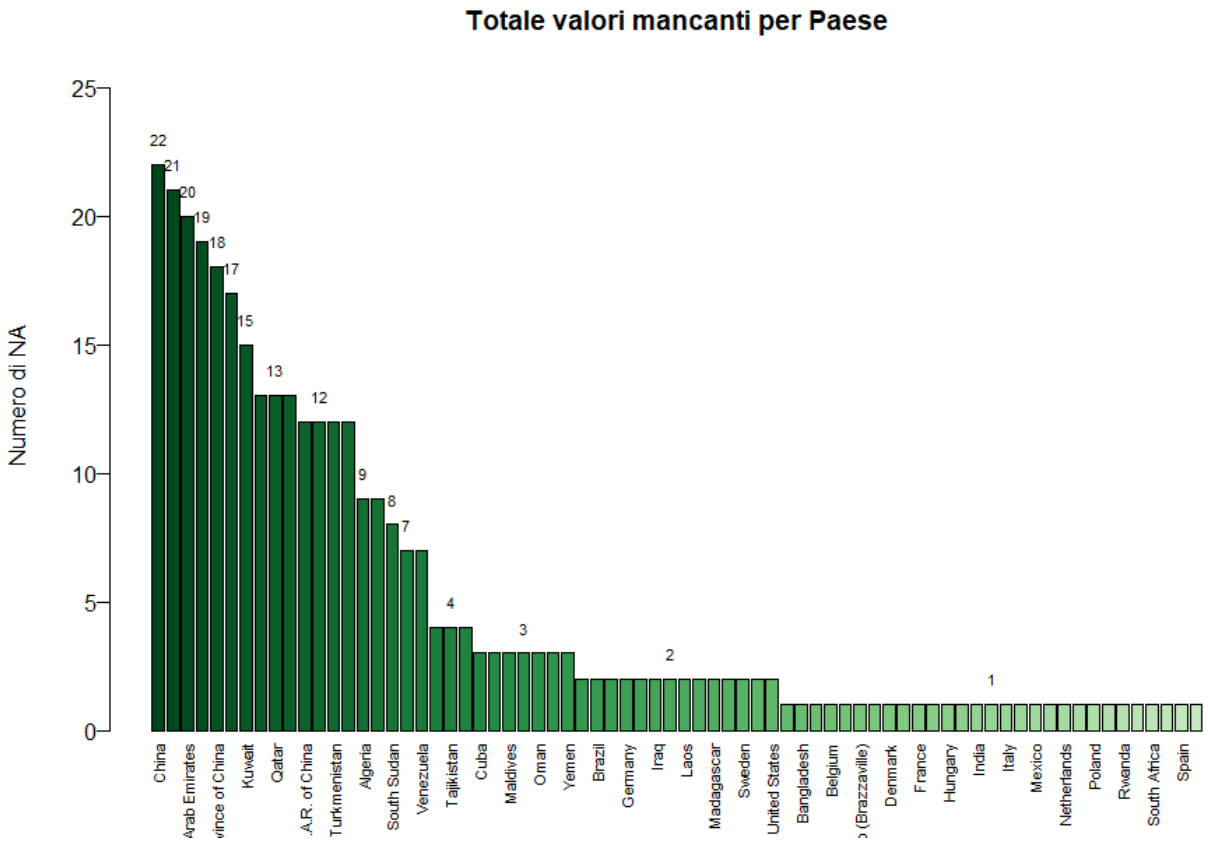


Figura 2.2: Barplot di distribuzione dei valori mancanti per ogni paese

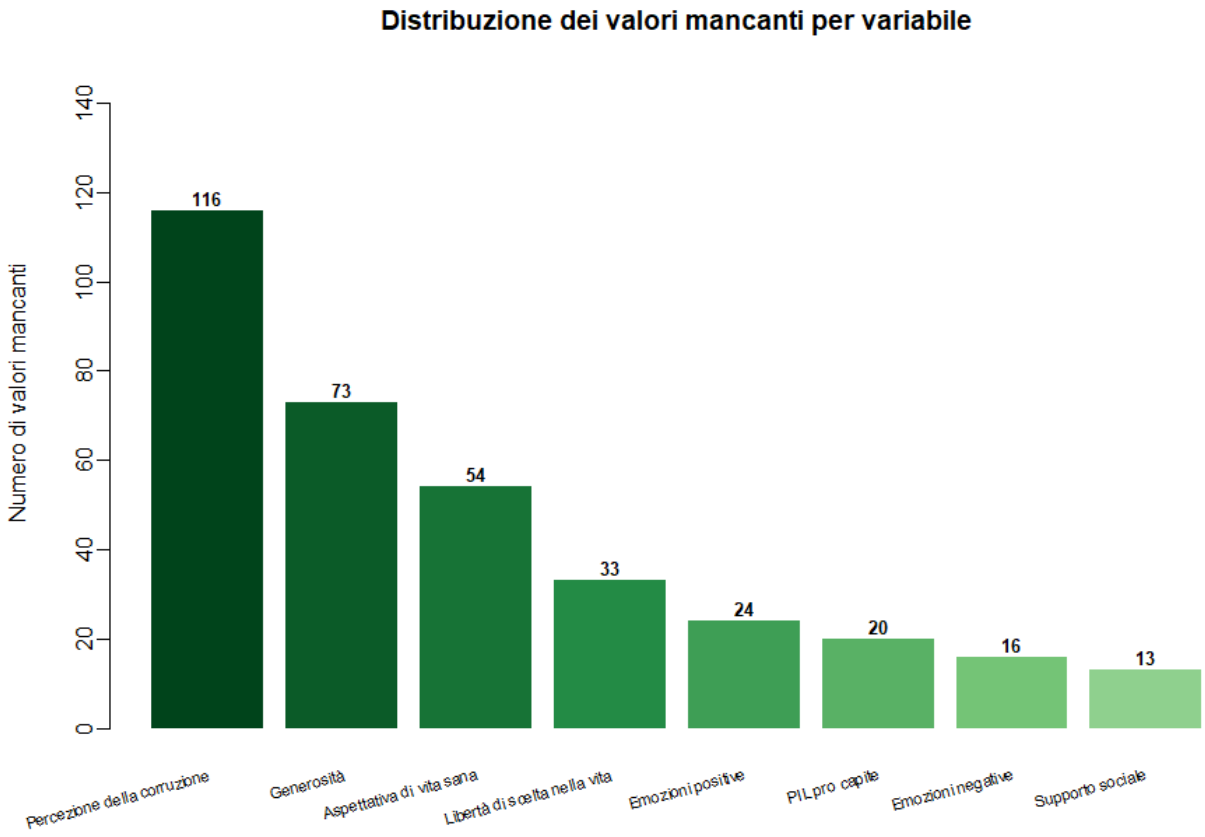


Figura 2.3: Barplot di distribuzione dei valori mancanti per ciascuna variabile del dataset

La mancanza di valori può introdurre bias nelle analisi statistiche, soprattutto in presenza di variabili socio–economiche correlate tra loro, e impedire la corretta analisi dei valori dei vari paesi. Per questo motivo, sono state valutate strategie di imputazione e applicati criteri di filtraggio:

- Esclusione dal dataset dei paesi che compaiono meno di due volte, in quanto la scarsità di osservazioni non consente una stima affidabile dei parametri e per permettere una distribuzione equa di dati per ogni anno;
- Imputazione statistica tramite la media del singolo paese per evitare l'influenza di paesi nettamente diversi.

Per garantire l'affidabilità dell'imputazione per paese, la media è stata utilizzata per sostituire i valori mancanti soltanto in presenza di almeno il 50% di valori validi fra gli anni considerati. Nei casi in cui tale soglia non fosse soddisfatta, la media è stata considerata non stimabile e il valore è stato lasciato mancante.

### 2.1.2 Valori Outlier

Per una migliore comprensione del dataset sono stati cercati gli eventuali outlier delle variabili (Figura 2.4).

Tramite l'osservazione dei boxplot si nota una presenza di outlier asimmetrici per la maggior parte delle variabili. Alcune, come le Emozioni negative e la Percezione della corruzione, ne presentano di più rispetto alle altre, mostrando una forte disparità fra alcuni paesi.

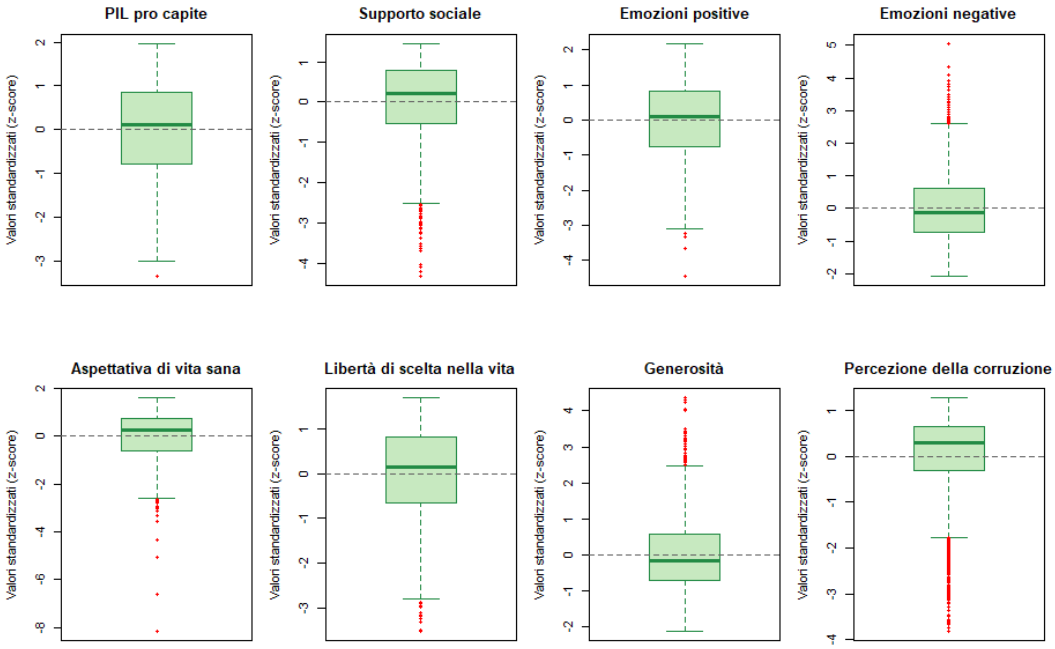


Figura 2.4: Boxplot outlier

### 2.1.3 Variabilità

Per comprendere la differenza dei punteggi di felicità tra i paesi considerati sono state calcolate le principali misure di dispersione statistica:

- La **varianza campionaria** associata al coefficiente di felicità è $s^2 = 1.2667$, indicando una dispersione moderata dei dati rispetto alla media, pur mostrando differenze significative tra gruppi di paesi;

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- La **deviazione standard campionaria** è $s = 1.1255$ e rappresenta lo scostamento dei valori rispetto alla media. La maggior parte dei paesi si colloca in una fascia di

benessere percepito medio-alta;

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Il **coefficiente di variazione**, con valore $CV = 20.53\%$, indica una variabilità moderata.

$$CV = \frac{s}{|\bar{x}|}$$

In conclusione, le misure di variabilità evidenziano una distribuzione della felicità relativamente omogenea e concentrata attorno ai valori medi.

### 2.1.4 Distribuzioni di frequenza

Per comprendere la struttura della felicità analizzata sono state osservate la frequenza assoluta e la frequenza relativa, raggruppando le osservazioni in classi, ognuna con uno scostamento di 0.15 rispetto alla precedente (Figure 2.5, 2.6).

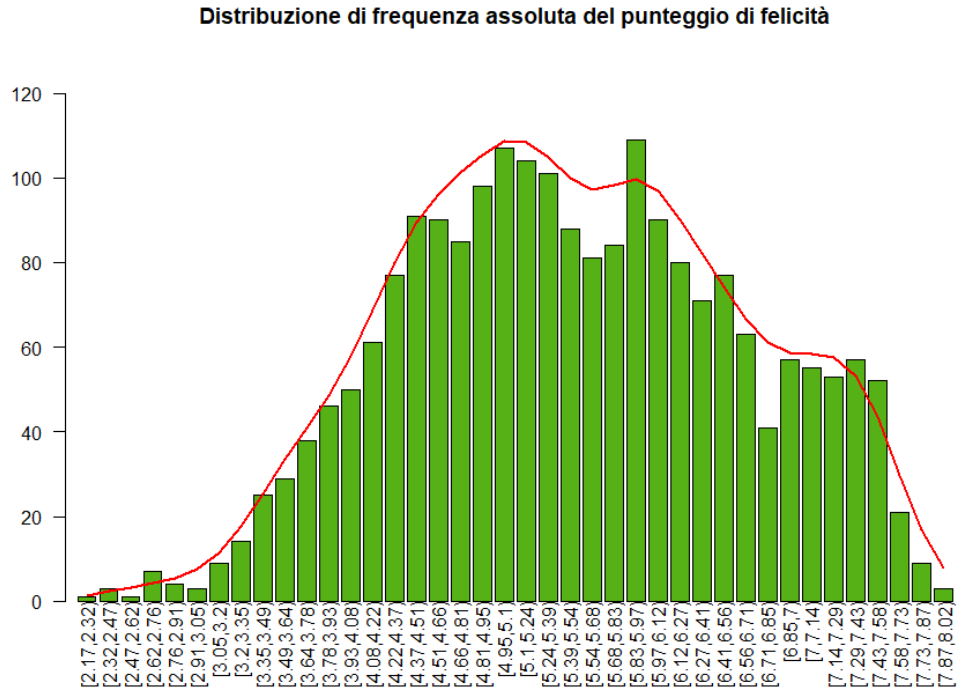


Figura 2.5: Distribuzione della frequenza assoluta di felicità

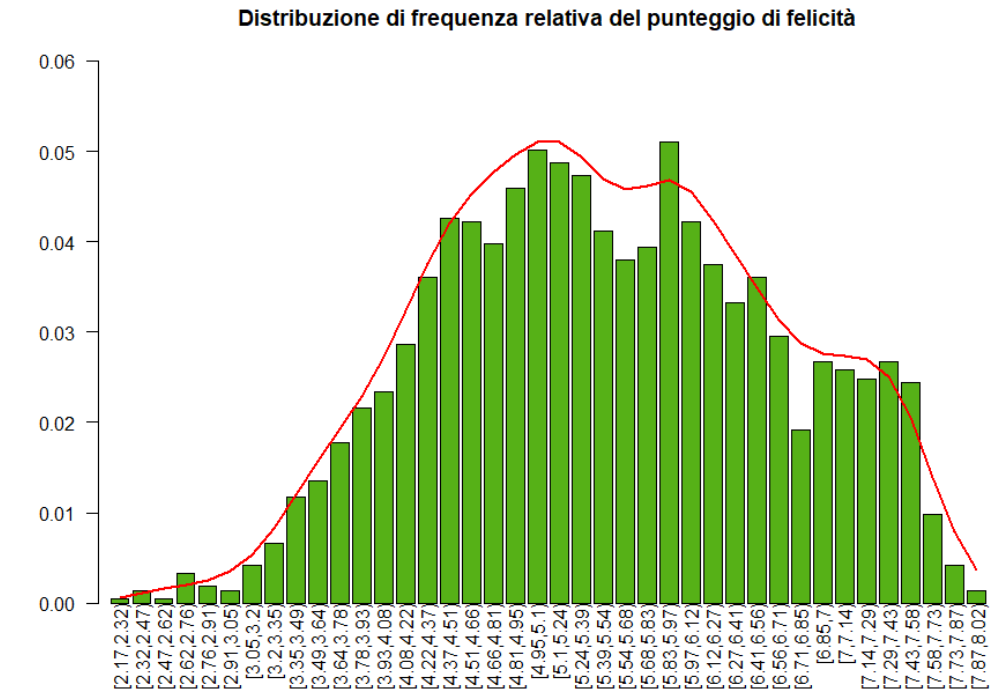


Figura 2.6: Distribuzione della frequenza relativa di felicità

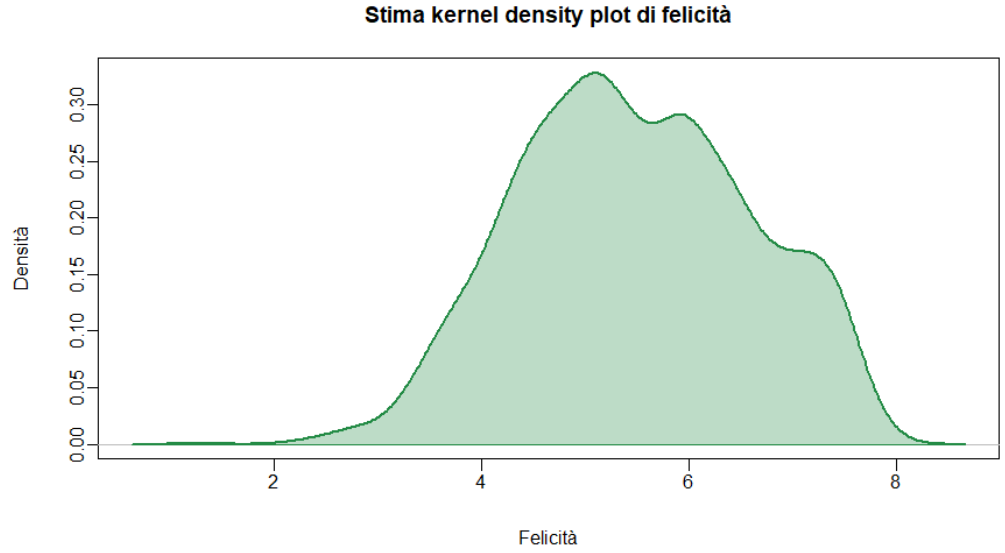


Figura 2.7: Kernel density plot della felicità

Il grafico della frequenza assoluta (Figura 2.5) mostra che la maggior parte delle osservazioni si concentra nelle classi centrali e non ci sono paesi con valore di felicità massimo (da 8 a 10) o minimo (da 0 a 2.17), come confermato dalle analisi numeriche (Tabella 2.5). La

distribuzione appare unimodale e mostra una decrescita graduale delle frequenze verso gli estremi. Ciò è confermato dalla frequenza relativa (Figura 2.6) e dalla stima kernel (Figura 2.7) che mostra una leggera asimmetria verso destra.

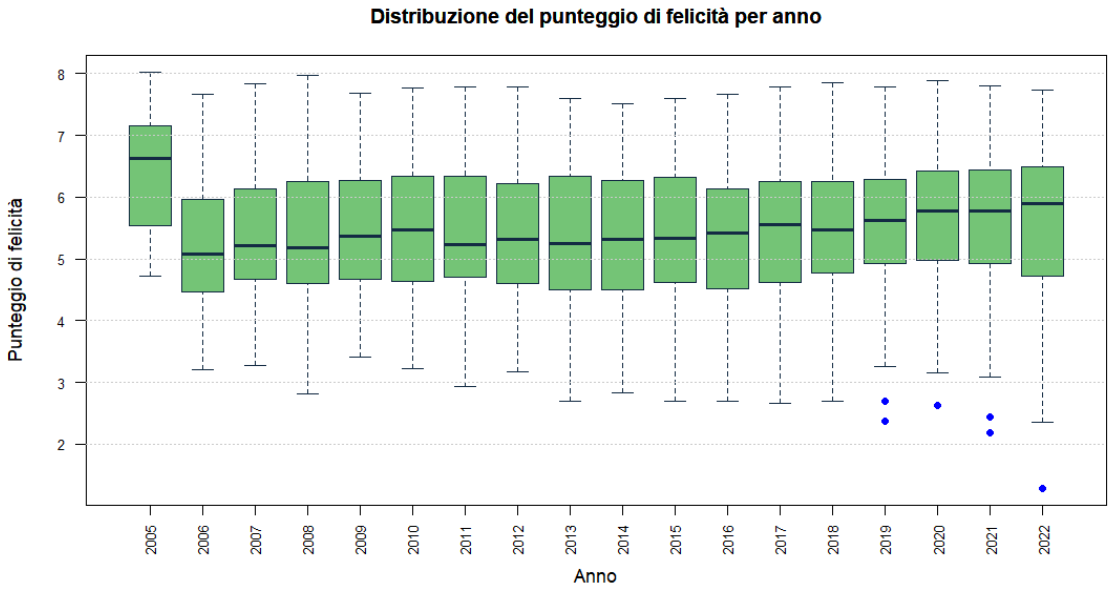


Figura 2.8: Boxplot di variazione del punteggio di felicità nel tempo

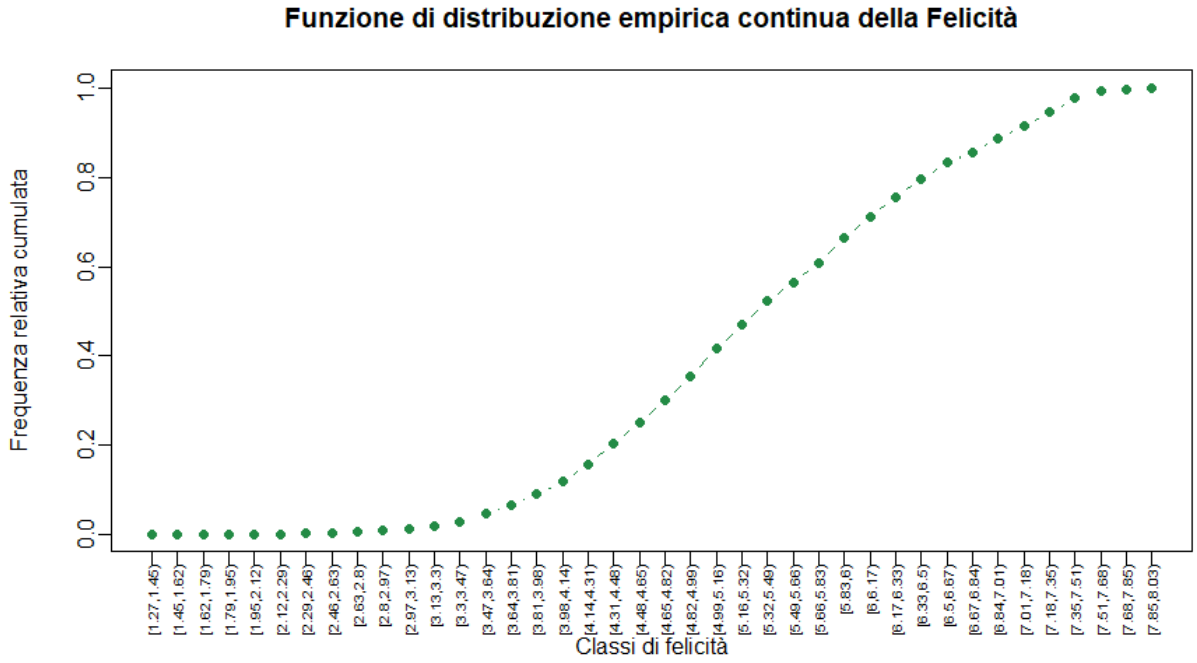


Figura 2.9: Funzione di distribuzione continua della felicità

Nella distribuzione della felicità vengono confermate le osservazioni precedenti, mostrando una distribuzione leggermente asimmetrica verso i valori più alti. Questo indica una tendenza generale verso livelli medi di benessere, ma con alcune nazioni che si distinguono per valori anomali (Figura 2.8). Inoltre, il tasso di felicità presenta una crescita graduale, come risulta dalla frequenza relativa cumulata (Figura 2.9).

| **Statistiche** | **Valore** |
|---|---|
| Minimo | 2.179 |
| 1° Quartile | 4.641 |
| Mediana | 5.435 |
| Media | 5.483 |
| 3° Quartile | 6.316 |
| Massimo | 8.019 |

Tabella 2.5: Statistiche descrittive e indici di sintesi della felicità.

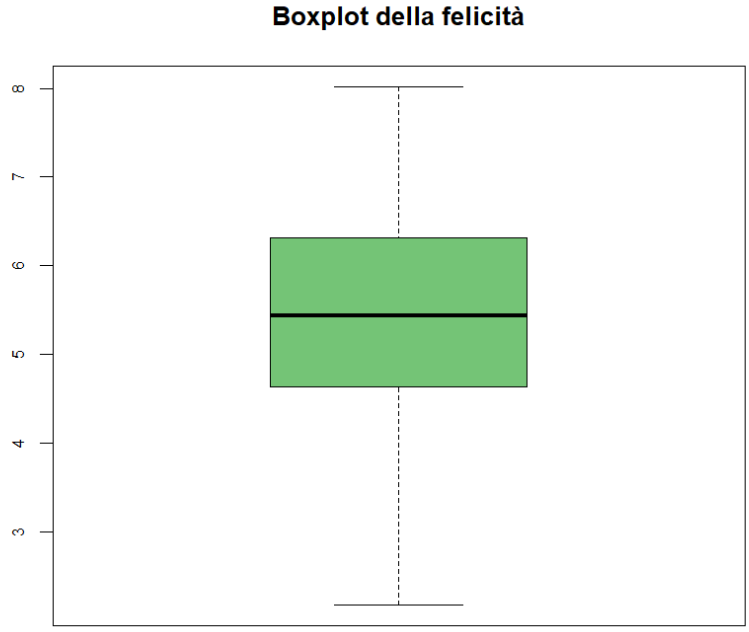


Figura 2.10: Boxplot del valore medio della felicità

Dall'analisi delle misure descrittive (Tabella 2.5) e dal boxplot sul valore medio di felicità (Figura 2.10) si osserva che la media è molto vicina alla mediana, evidenziando una distribuzione quasi simmetrica. L'intervallo interquartile ($IQR = 1.675$) indica una variabilità moderata.

La struttura della distribuzione mostra una leggera coda più estesa verso i valori inferiori con pochi punteggi molto bassi e distanti dalla media e una prevalenza di paesi con un alto valore di felicità. Tale comportamento può riflettere condizioni critiche presenti in alcuni paesi e suggerisce che la felicità, a livello internazionale, è influenzata principalmente da fattori socioeconomici condivisi.

### Simmetria

Per caratterizzare in modo più preciso la forma della distribuzione sono state svolte analisi sulla simmetria, sia graficamente che numericamente, tramite il coefficiente di simmetria (**skewness**).

$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} = \frac{m_3}{m_2^{3/2}}$$

Il valore globale ottenuto è pari a *0.0057*, molto vicino allo zero, indicando una distribuzione abbastanza simmetrica e bilanciata intorno alla mediana. Una skewness leggermente positiva descrive una leggera prevalenza di Paesi con punteggi di felicità lievemente superiori alla media rispetto a quelli collocati nella parte inferiore della distribuzione.

Complessivamente, i risultati indicano una distribuzione moderatamente simmetrica e priva di code pronunciate, confermata anche dal valore di **curtosi** (2.4068), caratteristica delle distribuzioni platicurtiche (una densità di probabilità leggermente più appiattita e con code meno pronunciate).

$$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 = \frac{m_4}{m_2^2} - 3$$

Ciò conferma le osservazioni precedenti.

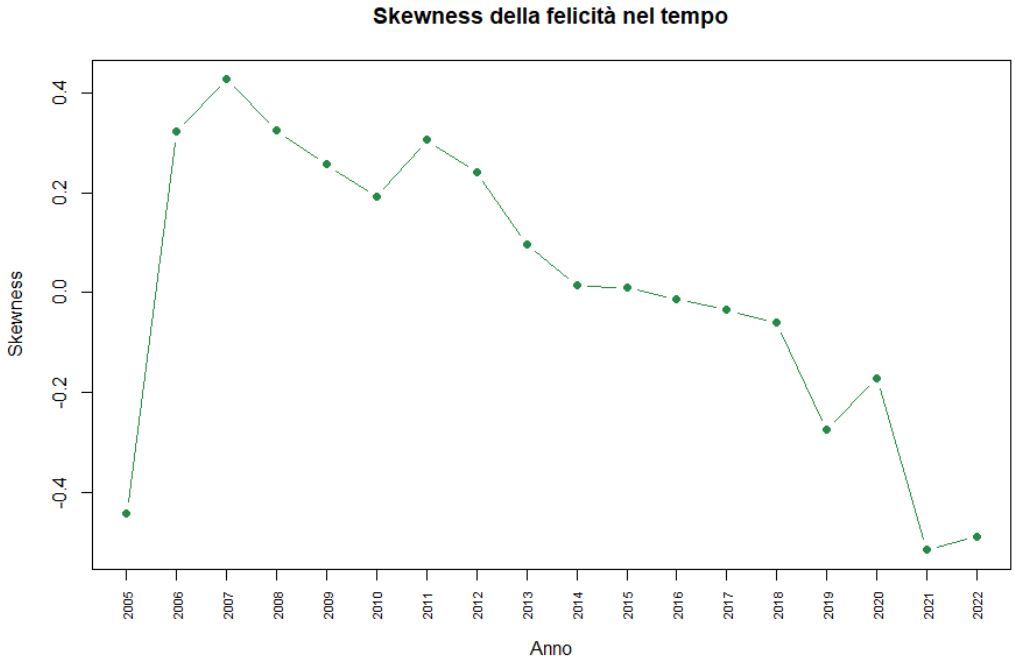


Figura 2.11: Grafico di variazione del valore della skewness per anno

L'analisi temporale della skewness calcolata per ogni anno (Figura 2.11) mostra come la forma della distribuzione abbia subito variazioni nel corso del periodo esaminato:

- Nel 2005 la skewness ha un valore negativo, seppure non il minimo rilevato. Ciò è dovuto anche a una mancanza di rilevazioni per molti paesi in quegli anni;
- Negli anni compresi tra il 2006 e il 2012 si osservano valori tendenzialmente positivi, con una maggiore concentrazione numerica di paesi con livelli di felicità superiori alla media.
- Dal 2019 la skewness assume valori progressivamente negativi;
- Nel 2021 si raggiunge la skewness minima, mostrando la presenza di molti valori inferiori rispetto alla media.

Ciò indica un progressivo spostamento della distribuzione verso punteggi medio-alti, con una riduzione relativa al periodo pandemico globale e al periodo di crisi successivo che hanno avuto un impatto socio-psicologico sui livelli di felicità rilevati.

## 2.2 Analisi grafiche

Per comprendere meglio la struttura del dataset sono state svolte delle analisi grafiche.

### 2.2.1 Analisi temporali

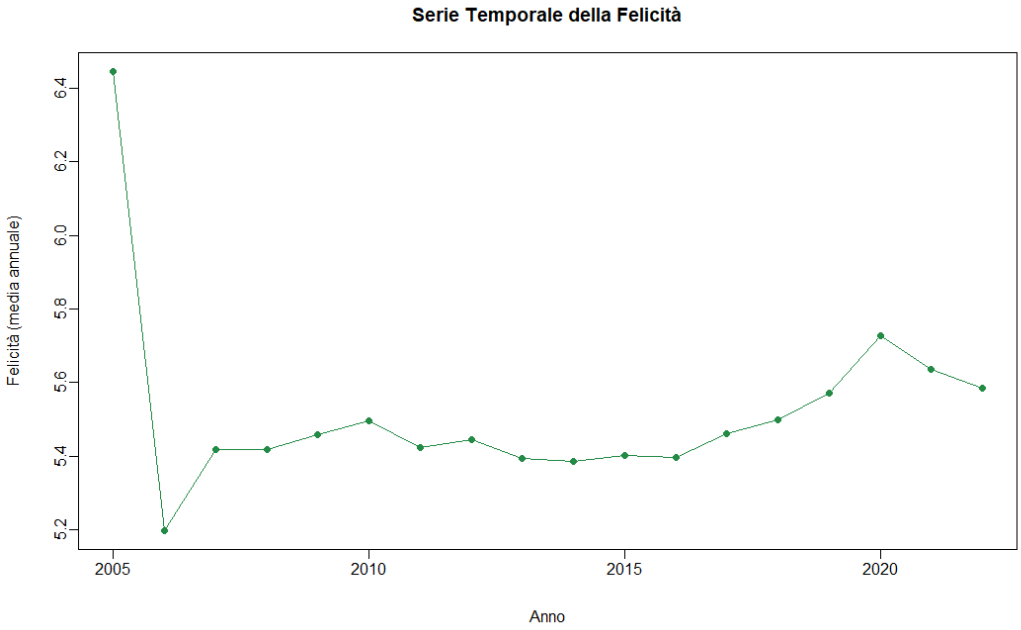


Figura 2.12: Serie temporale dal 2005 al 2022 sull'indice di felicità.

Per comprendere l'andamento della felicità è stata analizzata la serie temporale negli anni considerati (dal 2005 al 2022) (Figura 2.12):

- 2005: indice sensibilmente più alto rispetto agli anni successivi. Ciò può dipendere sia da fattori socioeconomici che generano un valore molto alto e anomalo sia dalla difficoltà di rilevazione per quell'anno;
- 2006: crollo visibile quasi sicuramente causato dall'entrata di più paesi con livelli di benessere medio-bassi;
- 2007-2015: felicità concentrata in valori medi, con una leggera ascesa verso la fine dell'intervallo. Ciò mostra condizioni socioeconomiche costanti;
- 2016-2019: crescita della felicità, probabilmente a causa della stabilizzazione post-crisi economica globale;
- 2020: picco relativo al periodo pandemico globale, dove gli effetti sociali del lockdown hanno fornito un aumento della percezione del supporto sociale e una riduzione della criminalità;
- 2021-2022: calo a causa di crisi e inflazione.

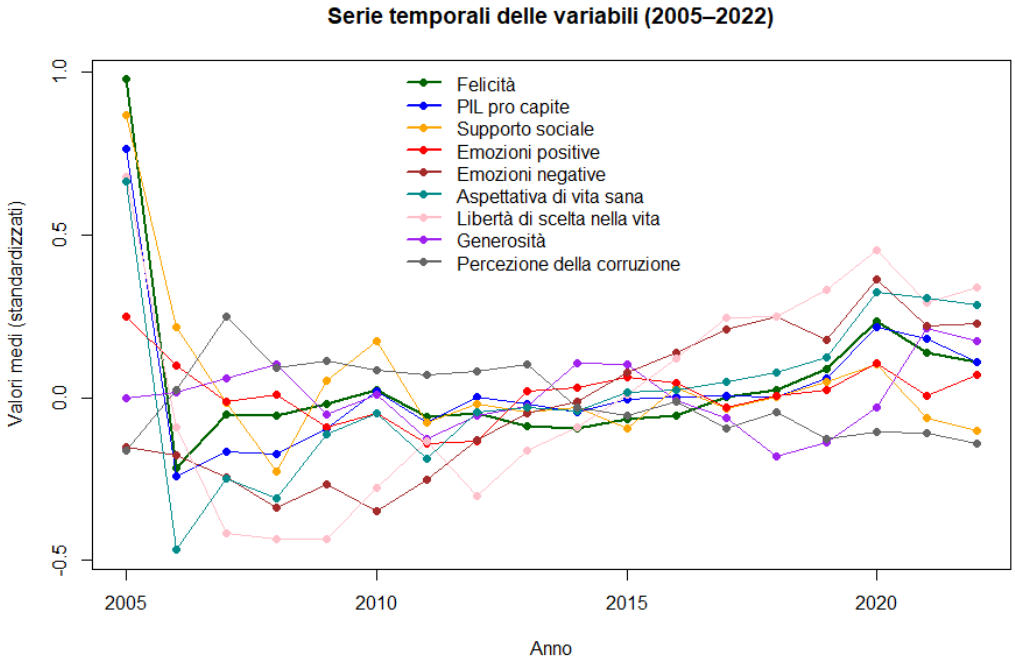


Figura 2.13: Serie temporale dal 2005 al 2022 sui valori.

È stato inoltre analizzato l'andamento temporale dei valori standardizzati delle variabili che influenzano la felicità per trovare eventuali correlazioni (Figura 2.13):

- 2005: picco della maggioranza dei valori probabilmente dovuto a mancanza di dati campionati;
- 2006: inizio della stabilizzazione dei valori con una sensibile discesa;
- 2007-2009: Picco positivo della Percezione di corruzione e negativo di Libertà di scelta nella vita e Supporto sociale, con le altre variabili che seguono percorsi graduali;
- 2010-2013: stabilizzazione e convergenza attorno allo 0 con qualche dato in salita;
- 2014-2016: convergenza intorno allo 0 della maggior parte dei valori;
- 2016-2020: aumento crescente di tutti i dati, ad eccezione di Generosità, che ha un picco in discesa;
- 2020-2022: piccola discesa da parte di quasi tutti i valori, probabilmente dovuto alla crisi post pandemia.

Nel complesso, le variabili presentano un andamento coerente fra loro con leggeri discostamenti, durante alcuni anni, di elementi che potrebbero risultare meno correlati agli altri.

### 2.2.2 Analisi della variazione delle variabili

Sono state analizzate, per ogni variabile, le scale dei paesi con il valore più alto e più basso per determinare se ci sono variazioni nette oppure se il dataset presenta variabili omogenee.

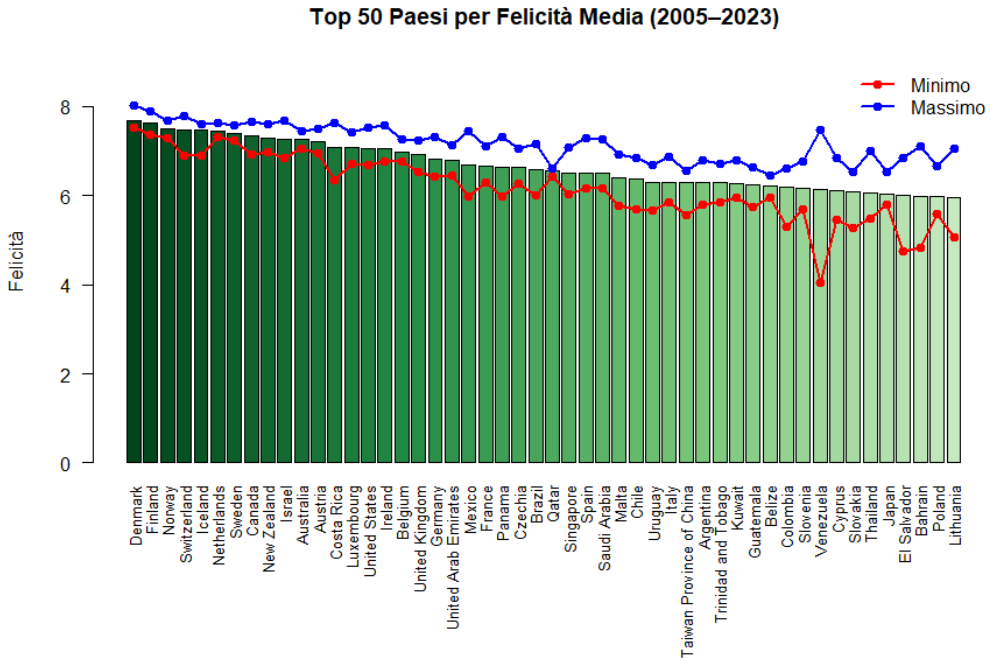


Figura 2.14: Top 50 Paesi con felicità media

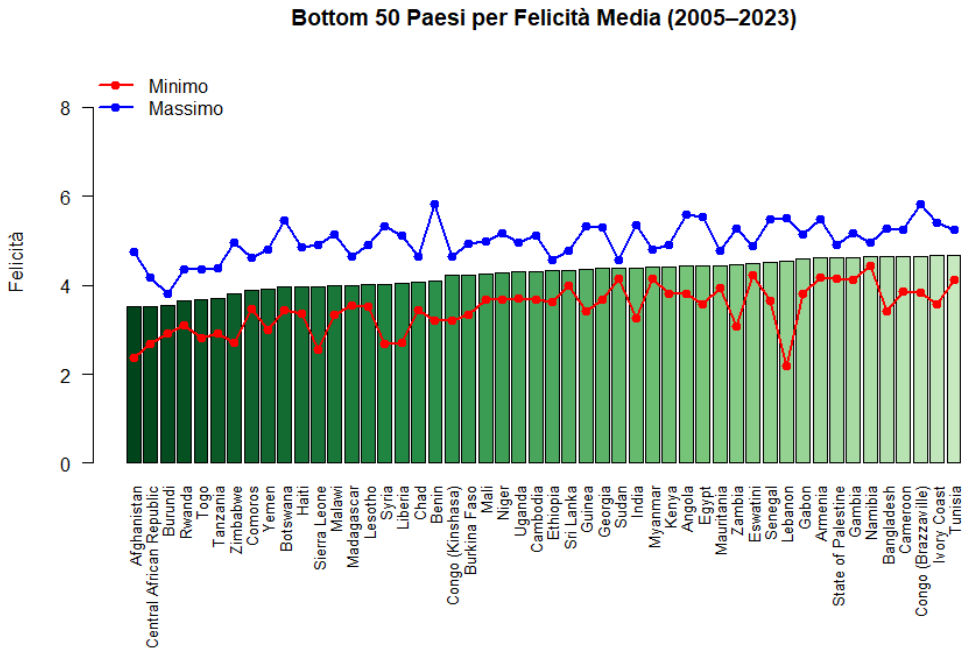


Figura 2.15: Bottom 50 Paesi con felicità media

Analizzando la felicità (Figure 2.14, 2.15) si nota come, nonostante ci siano nette variazioni fra il valore massimo e minimo per ogni paese, la media conserva una crescita costante, con paesi nettamente diversi che presentano una media di felicità simile.

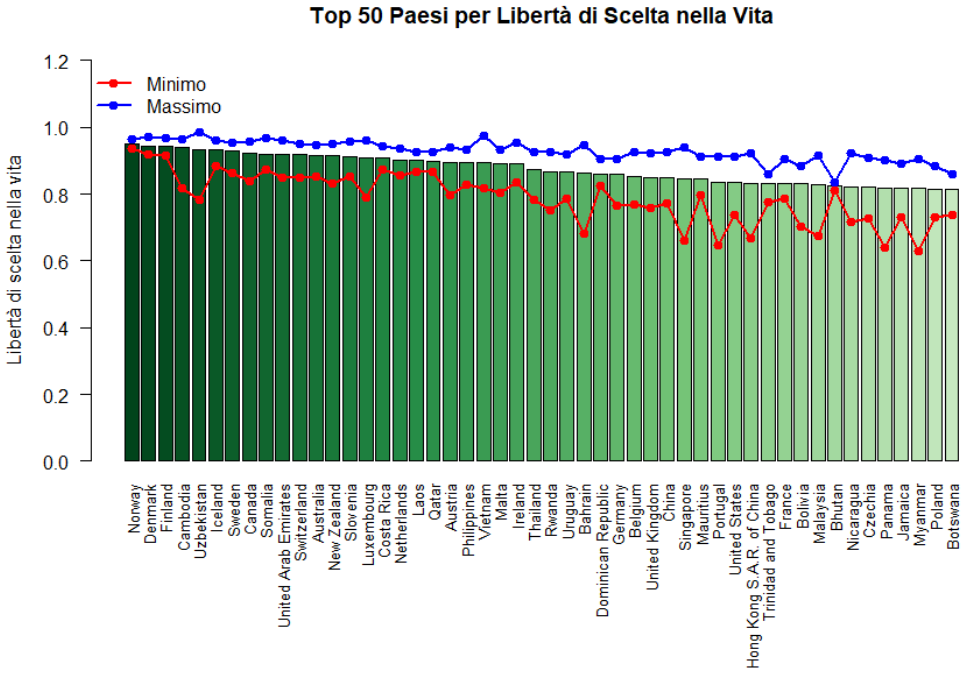


Figura 2.16: Top 50 Paesi con libertà di scelta media

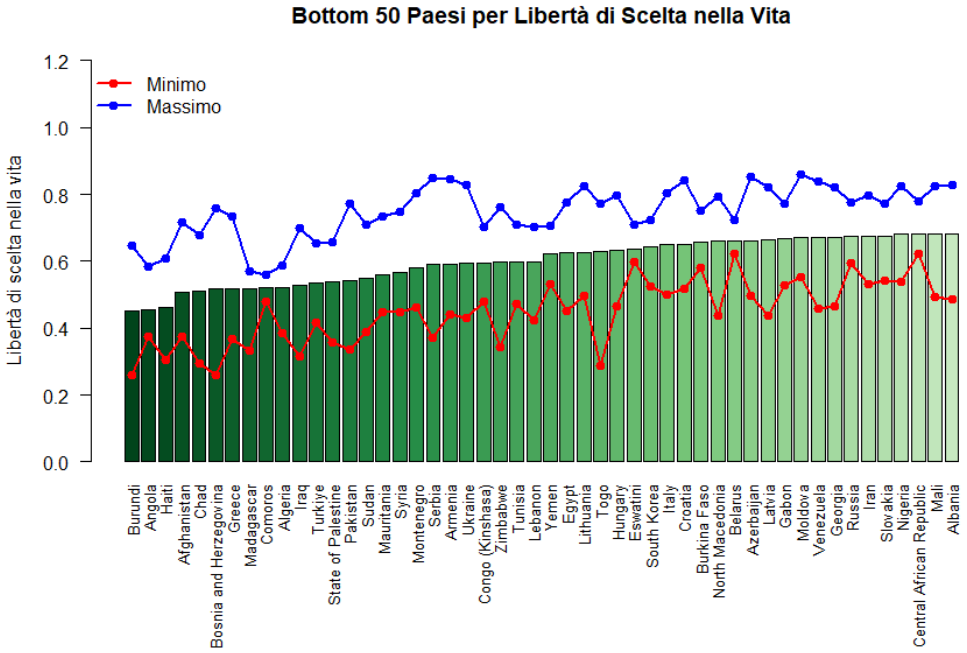


Figura 2.17: Bottom 50 Paesi con libertà di scelta media

Analizzando la libertà di scelta (Figure 2.16, 2.17) fra i primi paesi i valori non sono

molto discostati tra loro, mantenendo un massimo e minimo non troppo distaccati. Fra i paesi inferiori, invece, si rileva una forte differenza fra libertà massima e minima, e ciò può influenzare anche l'indice di felicità.

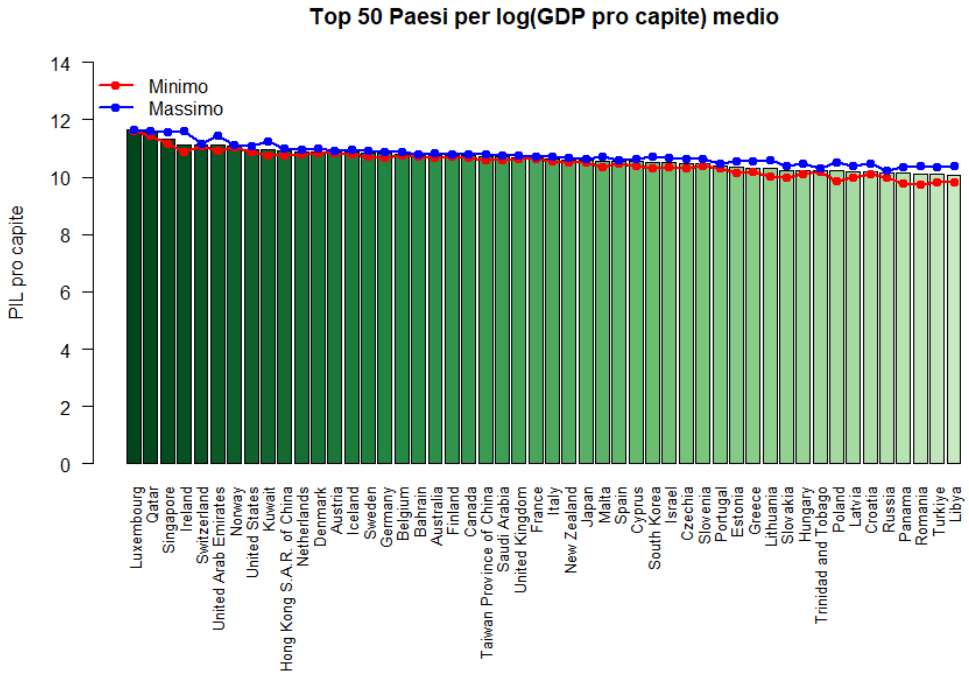


Figura 2.18: Top 50 Paesi con pro capite medio

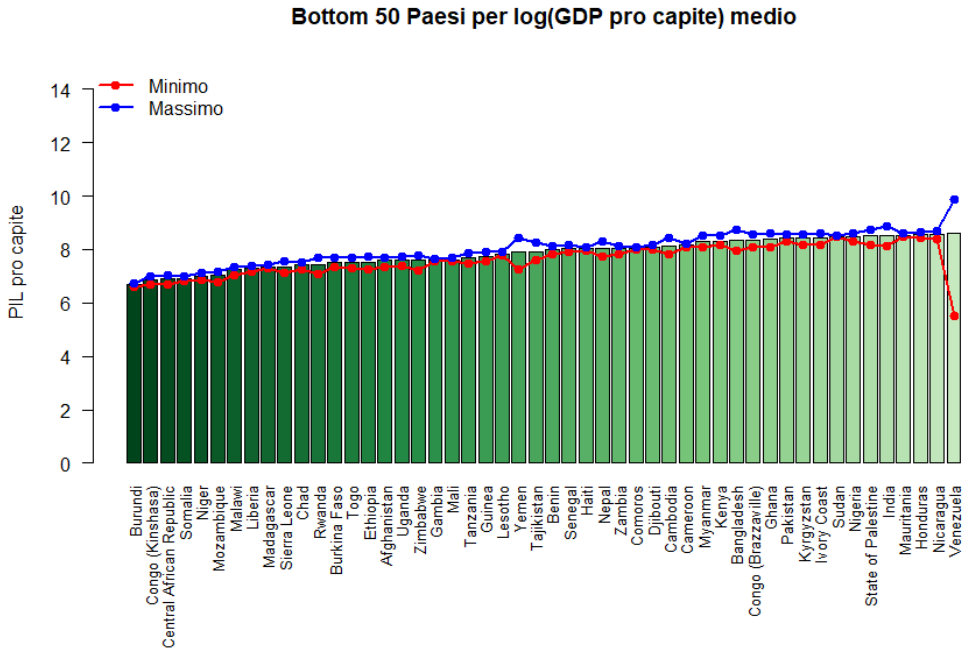


Figura 2.19: Bottom 50 Paesi con pro capite medio

Analizzando il pil pro capite (Figure 2.18, 2.19) si nota come i valori siano molto contenuti, con variazione fra minimo e massimo quasi nulla. L'unico paese con valori sbilanciati è il Venezuela.

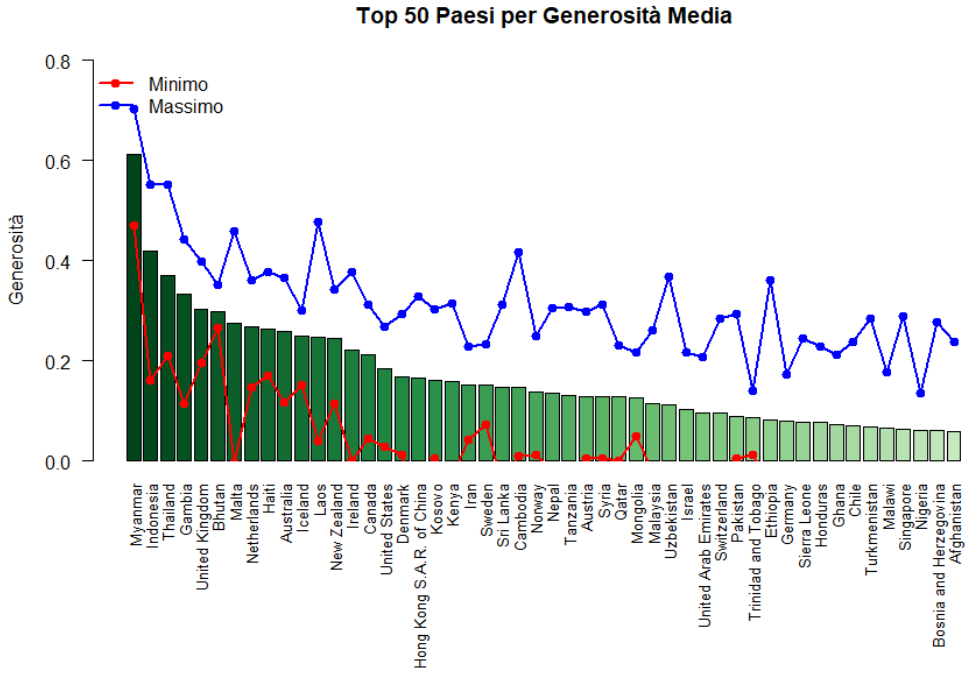


Figura 2.20: Top 50 Paesi con generosità media

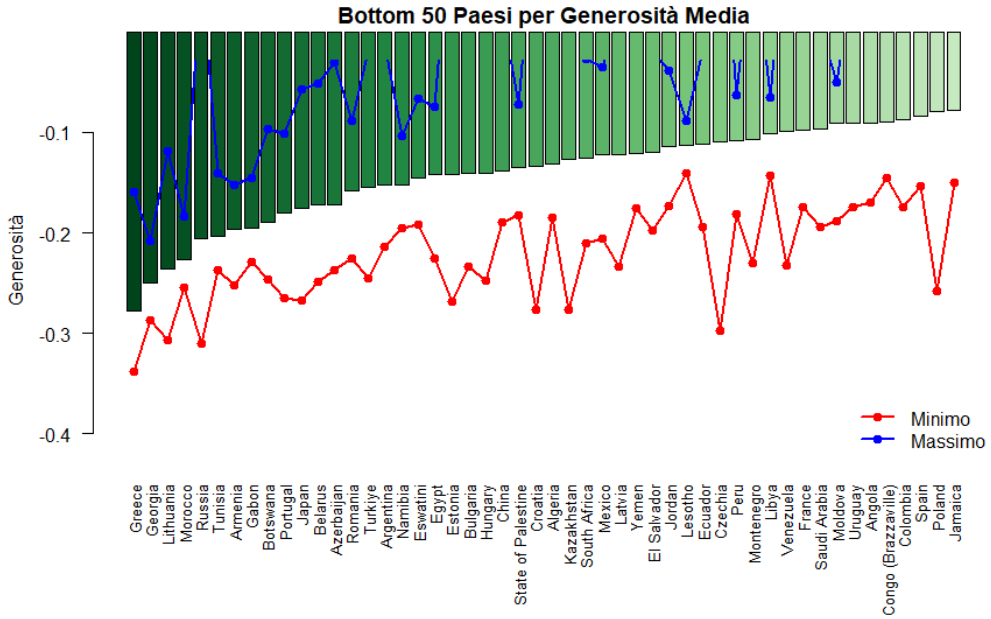


Figura 2.21: Bottom 50 Paesi con generosità media

Analizzando la generosità (Figure 2.20, 2.21) è evidente uno squilibrio fra i valori. Il

primo paese ha un valore estremamente alto mentre per gli altri hanno una curva discendente. Lo squilibrio tra i valori minimi e massimi è estremamente alto, rendendola una variabile complessa da valutare in modo affidabile. Gli ultimi paesi presentano poi valori estremamente bassi, discostandosi ampiamente dal massimo generale.

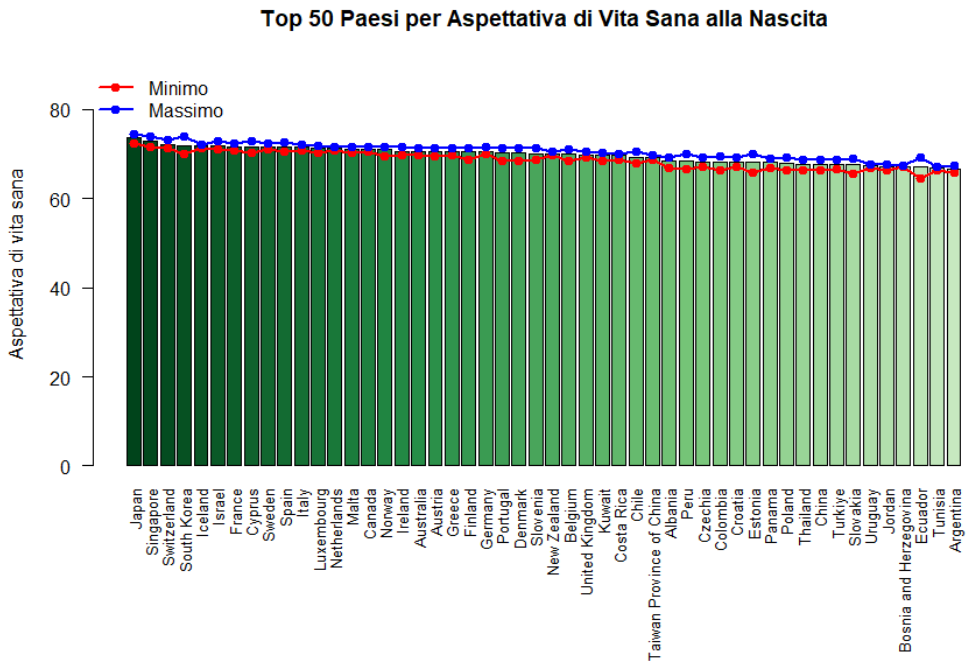


Figura 2.22: Top 50 Paesi con aspettative di vita alla nascita media

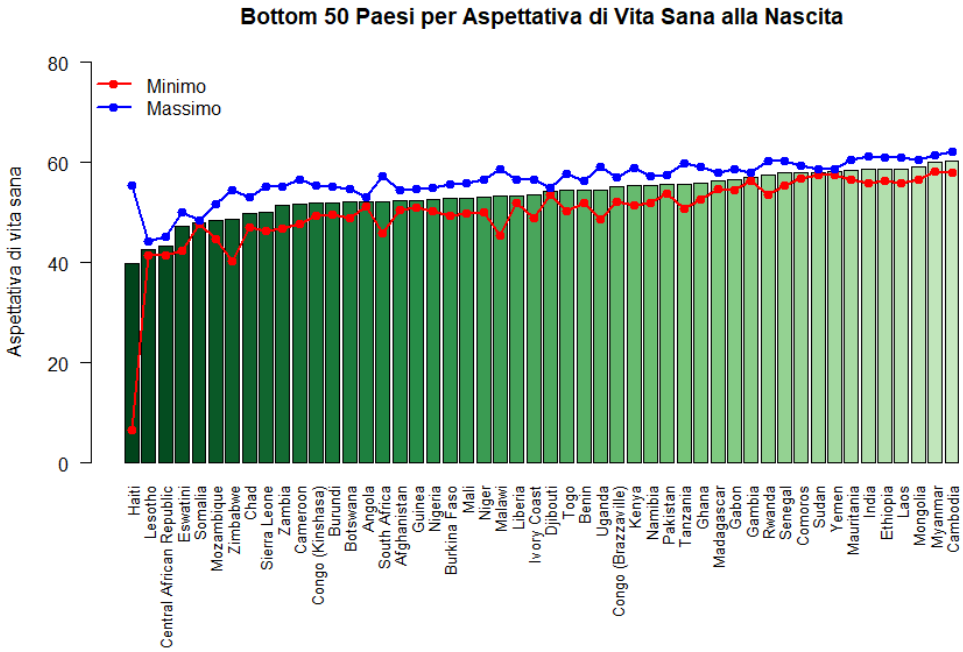


Figura 2.23: Bottom 50 Paesi con aspettativa di vita sana alla nascita media

Analizzando l'aspettativa di vita sana alla nascita (figure 2.22, 2.23) si nota un equilibrio fra i valori in quanto, soprattutto fra i primi paesi, la differenza fra il massimo e il minimo per ognuno è quasi nulla. Fra i paesi inferiori, invece, i valori per ognuno risultano più squilibrati, fino a raggiungere l'ultimo paese (Haiti) che presenta un massimo alto ma un minimo estremamente basso rispetto agli altri, risultando in una media distaccata dagli altri paesi.

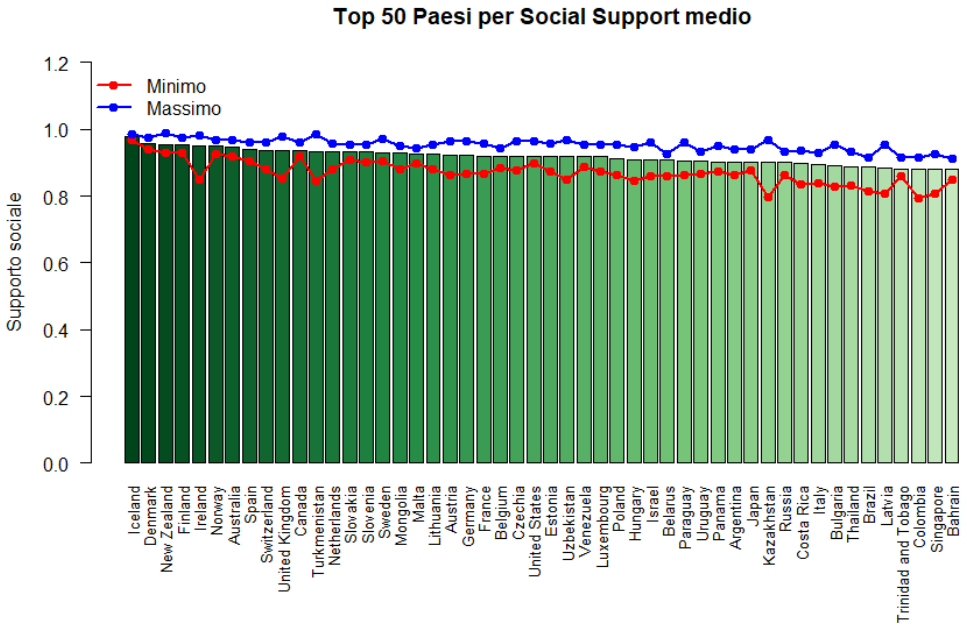


Figura 2.24: Top 50 Paesi con supporto sociale medio

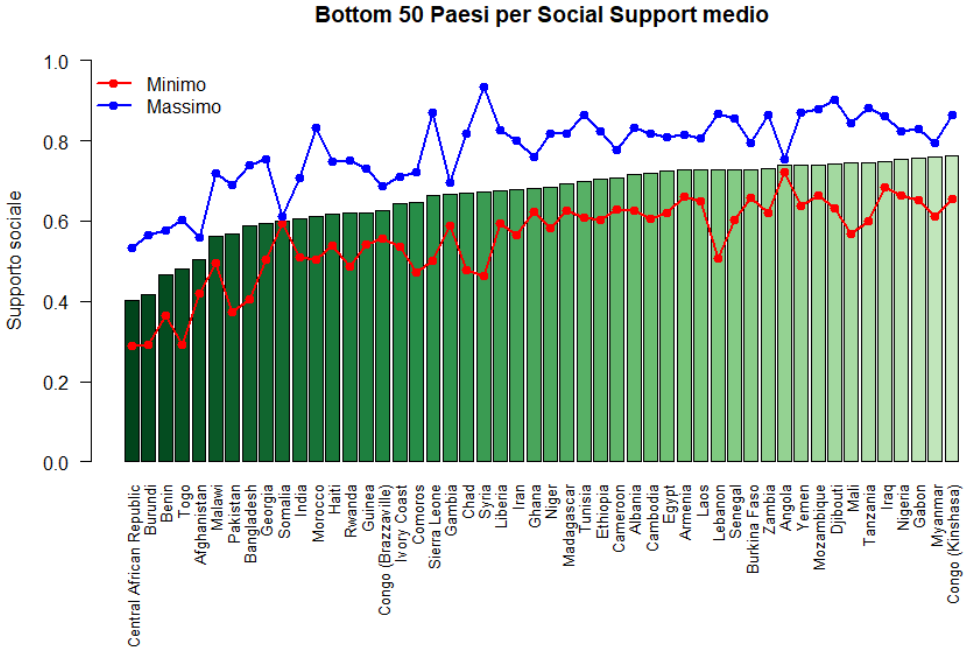


Figura 2.25: Bottom 50 Paesi con supporto sociale medio

Analizzando il supporto sociale (figure 2.24, 2.25) si nota un'equilibrio fra i valori più

alti ma uno squilibrio nei paesi più in basso. Infatti, i valori massimi e minimi sono distanti dalla media e i paesi in fondo alla scala presentano valori estremamente bassi e discostati dagli altri.

Le analisi mostrano come alcuni paesi presentano valori molto variegati. Le variabili con forti differenze fra top e bottom sono più importanti nel determinare la felicità rispetto a quelle che presentano variazioni minime.

### 2.2.3 Correlazioni tra variabili

Utilizzando un approccio di statistica descrittiva bivariata sono state messe a confronto le variabili con il punteggio di felicità per analizzare l'eventuale presenza di correlazioni.

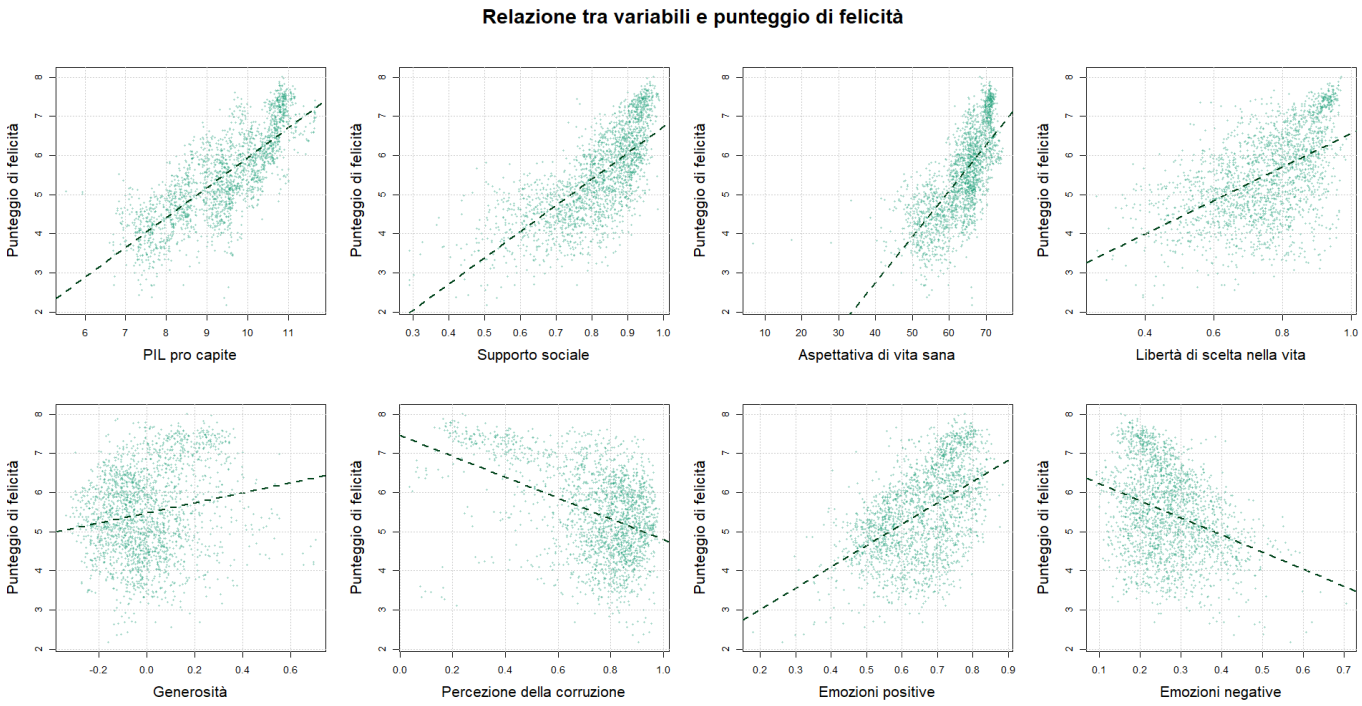


Figura 2.26: Scatterplot di relazione tra le variabili e il punteggio di felicità

Analizzando gli scatterplot (Figura 2.26) si osserva una correlazione positiva non perfettamente lineare fra Pil pro capite e punteggio di felicità: valori simili di ricchezza generano valori di felicità anche distanti dalla media, indicando la presenza di altri fattori che influenzano la felicità. L'economia quindi è un valore importante, seppure non l'unico, che genera un impatto nel determinare la felicità di un paese. La correlazione positiva emerge anche tra supporto sociale e punteggio di felicità: più è alto il supporto sociale maggiore è la felicità riscontrata.

Osservando l'aspettativa di vita sana alla nascita e il punteggio di felicità è evidente una correlazione positiva: i paesi con un'aspettativa di vita più elevata tendono ad avere anche un livello di felicità maggiore. Molti paesi, inoltre, si collocano tra i valori medio-alti di entrambi gli indicatori.
I paesi con un'aspettativa di vita più bassa mostrano, invece, punteggi di felicità inferiori, seppure la felicità non diminuisca linearmente con l'aspettativa di vita (paesi con aspettativa di vita fra 0 e 30 hanno un punteggio di felicità fra 3.5 e 4, quindi con una variazione minima, risultando come outlier). Ciò si nota anche osservando la correlazione fra generosità e punteggio di felicità: all'aumentare della generosità, tende ad aumentare anche il punteggio medio di felicità, tuttavia la dispersione ampia dei punti diminuisce l'impatto della variabile. Un andamento analogo si osserva per la libertà di scelta, dove nei paesi in cui gli individui si sentono più liberi di scegliere, il benessere soggettivo tende a essere maggiore.
La correlazione fra sentimenti positivi e punteggio di felicità denota una relazione positiva moderata. La linea di regressione lineare conferma una tendenza crescente, seppure ci sia una variabilità di punteggi di felicità relativa a vari gradi di sentimenti positivi. La variabile è quindi impattante sull'indice di felicità ma non l'unica da considerare.

Confrontando la relazione tra sentimenti negativi e punteggio di felicità si nota una correlazione negativa, come avviene osservando la percezione di corruzione: paesi con un'alta percezione di corruzione presentano un indice di felicità variabile, quindi non concentrato intorno alla media. Tuttavia sono presenti valori anomali, in quanto paesi con una bassa corruzione possono comunque essere infelici, così come quelli più felici possono comunque presentare un indice di corruzione alto.

Avendo trovato correlazioni tra le variabili e l'indice di felicità si è deciso di analizzarne i valori nel dettaglio successivamente (Sezione 3.1) tramite una domanda di ricerca (Sezione 2.3).

## 2.3 Domande di ricerca

Sulla base dell'analisi esplorativa effettuata precedentemente, sono state formulate le seguenti domande di ricerca:
*RQ1: Quali caratteristiche socioeconomiche influenzano maggiormente la felicità di un paese e in che misura tali relazioni causali possono essere validate attraverso test di inferenza?*
*RQ2: I dati generati da LLM (in particolare ChatGPT e Gemini) possono essere usati per migliorare la qualità del dataset reale? (es. integrando i dati per gli anni in cui non ci sono rilevazioni per determinati paesi)*
*In particolare, quale livello di coerenza statistica e strutturale tali dati raggiungono rispetto*

*al dataset reale, considerando la distribuzione delle feature, la valutazione delle principali tecniche di imputazione e i potenziali rischi di data leak.*

# Research question 1

***Quali caratteristiche socioeconomiche influenzano maggiormente la felicità di un paese e in che misura tali relazioni causali possono essere validate attraverso test di inferenza?***

Analizzando gli scatterplot tramite l'analisi grafica (sezione 2.2) si è riscontrata un'influenza delle variabili sull'indice di felicità. La risposta alla domanda è stata fornita, dopo le osservazioni grafiche, sia tramite analisi numeriche sulla dipendenza fra le variabili che tramite test di inferenza.

## 3.1 Covarianza e correlazione

| **Variabile** | **Correlazione** | **Segno della relazione** |
|---|---|---|
| PIL pro capite | 0.785 | Positiva |
| Supporto sociale | 0.720 | Positiva |
| Aspettativa di vita sana | 0.718 | Positiva |
| Libertà di scelta nella vita | 0.530 | Positiva |
| Emozioni positive | 0.515 | Positiva |
| Generosità | 0.183 | Positiva |
| Percezione della corruzione | -0.439 | Negativa |
| Emozioni negative | -0.333 | Negativa |

Tabella 3.1: Correlazione tra le variabili e il punteggio di felicità

$$r_{XY} = \frac{\text{Covarianza}(X,Y)}{\text{Scarti della media campionaria}} = \frac{\text{Cov}(X,Y)}{s_X\, s_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Poiché le variabili considerate sono espresse in unità di misura differenti è stato usato l'indice di correlazione campionario (Tabella 3.1) che determina una misura quantitativa della correlazione tra le variabili:

- Il PIL pro capite mostra la correlazione più elevata. Ciò suggerisce che paesi più ricchi tendono ad avere valori di felicità maggiori, confermando un legame diretto tra benessere economico e qualità percepita della vita;
- Il supporto sociale e l'aspettativa di vita sana presentano correlazioni positive relativamente elevate con la felicità. Ciò indica che la felicità è influenzata fortemente anche da fattori sociali, oltre che economici;
- La libertà di scelta nella vita e le emozioni positive mostrano relazioni moderate ma significative, indicando l'importanza di libertà individuale e stati emotivi positivi;
- La generosità presenta una lieve correlazione positiva, suggerendone un ruolo secondario;
- La percezione della corruzione e le emozioni negative risultano invece negativamente correlate con la felicità.

Il diagramma di Pareto è stato utilizzato per rappresentare l'importanza delle variabili nell'influenzare il punteggio di felicità. Il grafico (Figura 3.1) evidenzia che un numero ristretto di variabili (PIL pro capite, supporto sociale e aspettativa di vita) spiega la variabilità del punteggio di felicità meglio rispetto alle altre che, tuttavia, hanno comunque un'influenza importante.

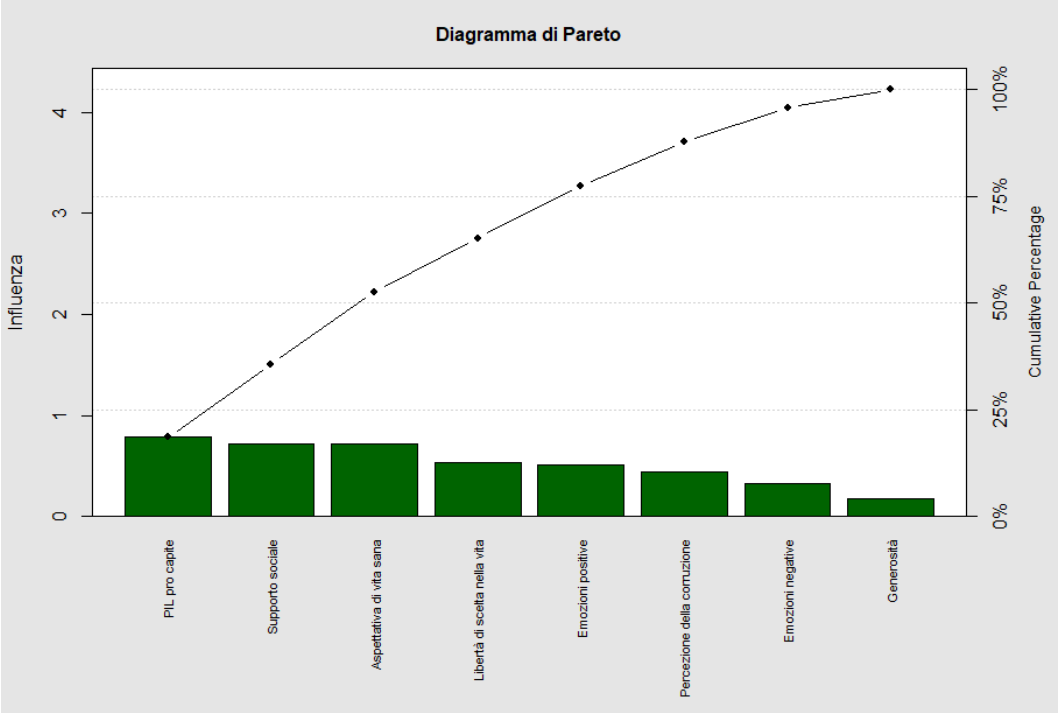


Figura 3.1: Diagramma di Pareto.

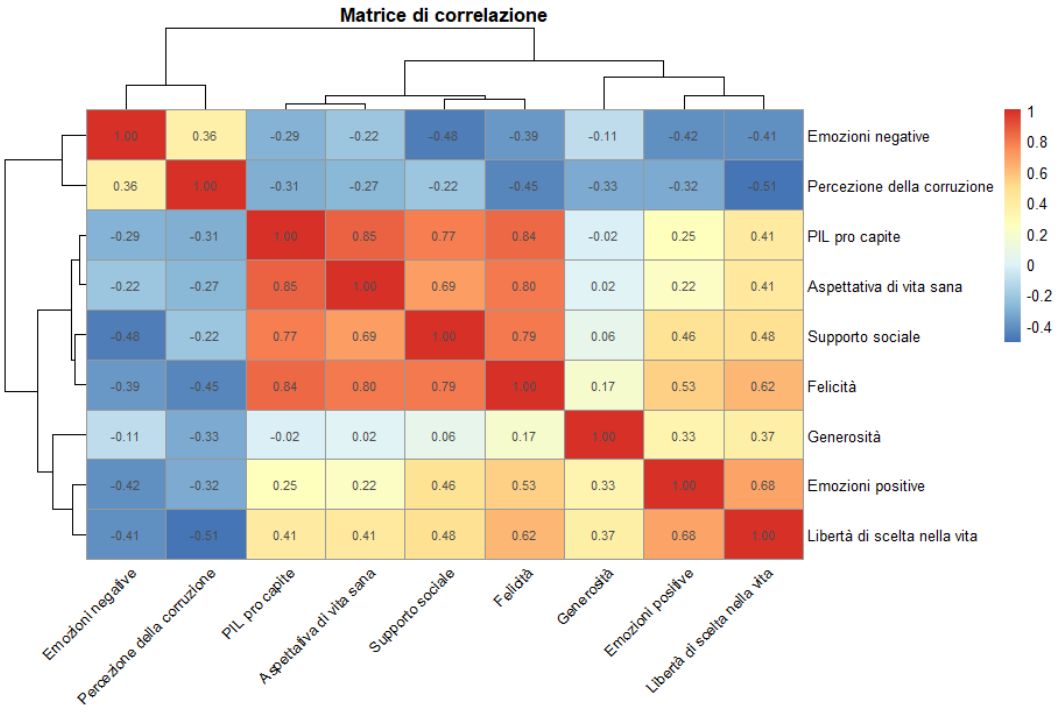


Figura 3.2: Matrice di correlazione fra le variabili

Per rappresentare graficamente le relazioni tra le variabili è stata usata la matrice di correlazione (Figura 3.2). Si nota una forte correlazione positiva tra il PIL pro capite, il supporto sociale e l'aspettativa di vita in buona salute alla nascita, suggerendo che i paesi economicamente più sviluppati tendono ad avere anche migliori condizioni sociali e sanitarie.

La correlazione dei sentimenti negativi e della percezione di corruzione, invece, risulta fortemente negativa con le variabili positive. Fra loro le variabili presentano una lieve correlazione.

Viene evidenziato come l'indice di felicità dipenda da un insieme di elementi interconnessi, fra cui il supporto sociale, la ricchezza e l'aspettativa di vita e, in negativo, dalle variabili evidenziate precedentemente. La variabile meno impattante risulta essere la generosità.

## 3.2 Regressione Lineare

### 3.2.1 Regressione lineare semplice

Tenendo conto della correlazione evidenziata dagli scatterplot tra il punteggio di felicità e le principali variabili socioeconomiche (Sezione 2.2.3, Tabella 3.2), è stata condotta un'analisi esplorativa basata su regressione lineare semplice analizzando i residui, ovvero la differenza tra i valori osservati del punteggio di felicità e quelli stimati dal modello per rappresentare

l'errore commesso nella previsione del punteggio e controllare la presenza di pattern non lineari.

I valroi della retta interpolante mostrano una forte variazione della felicità al variare del supporto sociale, della libertà di scelta nella vita e delle emozioni positive, con una variazione negativa a causa della percezione della corruzione e delle emozioni positive. Le altre variabili provocano una variazione positiva seppure non fortemente impattante, come osservato graficamente negli scatterplot.

| **Variabile** | **Intercetta ($\alpha$)** | **Coefficiente angolare ($\beta$)** | **Segno Atteso** |
|---|---|---|---|
| PIL pro capite | -1.692 | 0.764 | Positivo |
| Supporto sociale | 0.040 | 6.711 | Positivo |
| Aspettativa di vita sana | -1.945 | 0.117 | Positivo |
| Libertà di scelta nella vita | 2.281 | 4.277 | Positivo |
| Generosità | 5.482 | 1.279 | Positivo |
| Percezione della corruzione | 7.468 | -2.680 | Negativo |
| Emozioni positive | 1.923 | 5.459 | Positivo |
| Emozioni negative | 6.673 | -4.379 | Negativo |

Tabella 3.2: Coefficiente lineare del confronto di ciascun attributo con la felicità.

I residui permettono di determinare la differenza fra la felicità osservata e quella prevista dal modello. Dall'osservazione dei grafici dei residui emerge un discostameto fra valori osservati e valori stimati non emergono pattern strutturati o curvature evidenti, in quanto risultano dispersi in modo casuale attorno allo zero. Sono presenti alcuni outlier senza però un impatto rilevante.

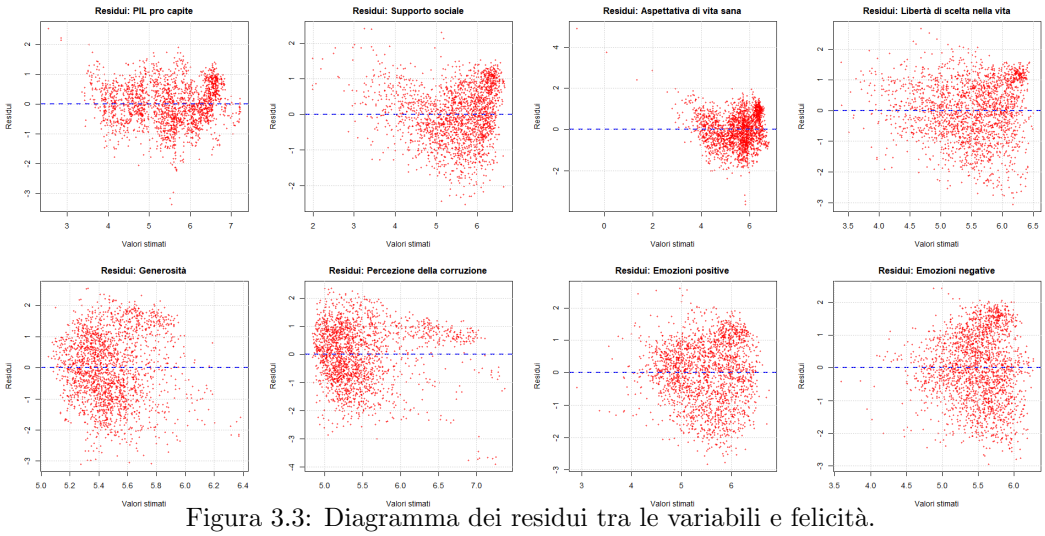


Figura 3.3: Diagramma dei residui tra le variabili e felicità.

- Osservando i residui relativi al Pil pro capite si nota una dispersione omogenea con una densità più elevata nei valori stimati alti.

- Nel diagramma dei residui relativo alla generosità è possibile notare una densità concentrata nella parte destra con valori dispersi e residui molto bassi.
- I residui relativi al supporto sociale risultano irregolari con una dispersione non del tutto uniforme lungo il dominio dei valori stimati: i valori più bassi sono dispersi con residui positivi e meno numerosi, mentre quelli più elevati si osserva un addensamento marcato di punti.
- Nel diagramma dei residui relativo alla percezione di corruzione è possibile notare una densità nei valori stimati a sinistra con valori dispersi. Dalla parte centrale alla sezione di destra si nota una bassa densità con una prevalenza di residui positivi.
- Osservando i residui dell'aspettativa di vita sana si nota una tendenza ad associare valori stimati più bassi a residui positivi e quelli più alti a negativi.
- Questo andamento dice che che il modello non riesce a catturare completamente la relazione tra le variabili, non essendo lineare.
- Nel diagramma dei residui relativo alle emozioni positive la dispersione dei punti risulta elevata ma con valori stimati medio-alti.
- I residui relativi alla libertà di scelta risultano distribuiti in modo casuale attorno allo zero. La dispersione non è completamente omogenea lungo l'asse: per valori stimati più elevati si osserva una maggiore concentrazione, mentre per valori più bassi i residui risultano più dispersi.
- Nel diagramma dei residui relativo ai sentimenti negativi si osserva che i punti risultano distribuiti in modo sostanzialmente casuale attorno alla linea orizzontale dello zero, senza evidenti strutture o tendenze. Ciò suggerisce che è presente una linearità tra la variabile e il punteggio di felicità, nonostante la dispersione dei residui sia lievemente asimmetrica.

### 3.2.2 Regressione Lineare multipla

Sulla base delle correlazioni evidenziate tramite la regressione lineare semplice, è stato stimato un modello di regressione lineare multipla, utilizzando la felicità come variabile dipendente e le altre variabili come regressori.

| Variabile | Coef. stimato ($\beta$) | Errore standard | Interv. confid. |
|---|---|---|---|
| Intercetta | 5.47513 | 0.03321 | [5.4100, 5.5402] |
| PIL pro capite | 0.39043 | 0.06928 | [0.2547, 0.5262] |
| Supporto sociale | 0.18136 | 0.06373 | [0.0565, 0.3063] |
| Aspettativa di vita sana | 0.26528 | 0.06011 | [0.1475, 0.3831] |
| Libertà di scelta nella vita | 0.07262 | 0.06082 | [-0.0466, 0.1918] |
| Generosità | 0.04232 | 0.04211 | [-0.0402, 0.1248] |
| Percezione della corruzione | -0.12628 | 0.04392 | [-0.2123, -0.0403] |
| Emozioni positive | 0.26753 | 0.05347 | [0.1627, 0.3723] |
| Emozioni negative | 0.02389 | 0.04942 | [-0.0730, 0.1207] |
| **Coefficiente di determinazione:** 0.863 | | | |

Tabella 3.3: Risultati della regressione lineare multipla sulla felicità media per Paese.

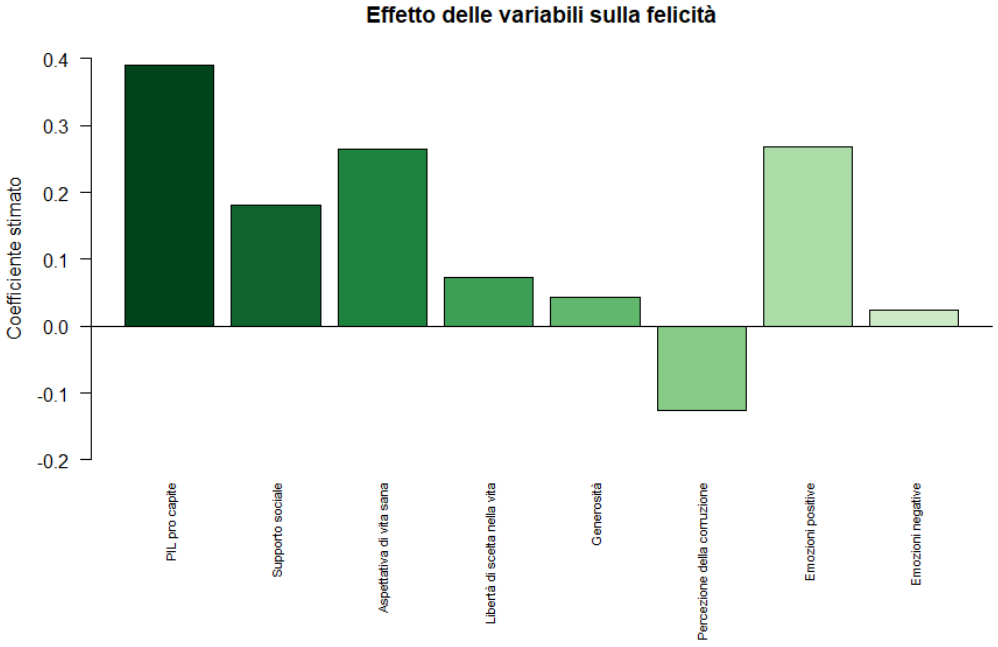


Figura 3.4: Regressione lineare della felicità.

Il punteggio medio di felicità dei Paesi è fortemente influenzato da fattori economici, sociali e psicologici determinati tramite il coefficiente stimato (che rappresenta di quanto varia il punteggio di felicità all'aumentare di un'unità la variabile) (Tabella 3.3). L'intercetta mostra il valore medio della felicità (quando non è influenzata dalle altre variabili), e il coefficiente stimato mostra quindi di quanto aumenta la felicità all'aumentare di 1 unità di goni variabile. L'errore standard molto basso conferma una stima precisa. In particolare, le variabili significative risultano essere: il PIL pro capite ($\beta = 0.39$), il supporto sociale ($\beta = 0.18$), l'aspettativa di vita in buona salute ($\beta = 0.26$) e le emozioni positive ($\beta = 0.27$).

(Figura 3.4) La percezione della corruzione mostra invece un effetto negativo significativo ($\beta = -0.13$).

Ciò è confermato dagli intervalli di confidenza che non comprendono lo 0, mostrando invece un impatto non significativo di libertà di scelta, emozioni negative e generosità.

Il coefficiente di determinazione (dato dal rapporto della varianza dei valori stiati con la retta di regressione e la varianza dei valori osservati)

$$D^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

indica che i regressori forniscono una spiegazione robusta della variabilità dell'indice di felicità e solamente il 14% è rappresentato dai residui (Figura 3.5) (non spiegata dalle variabili considerate).

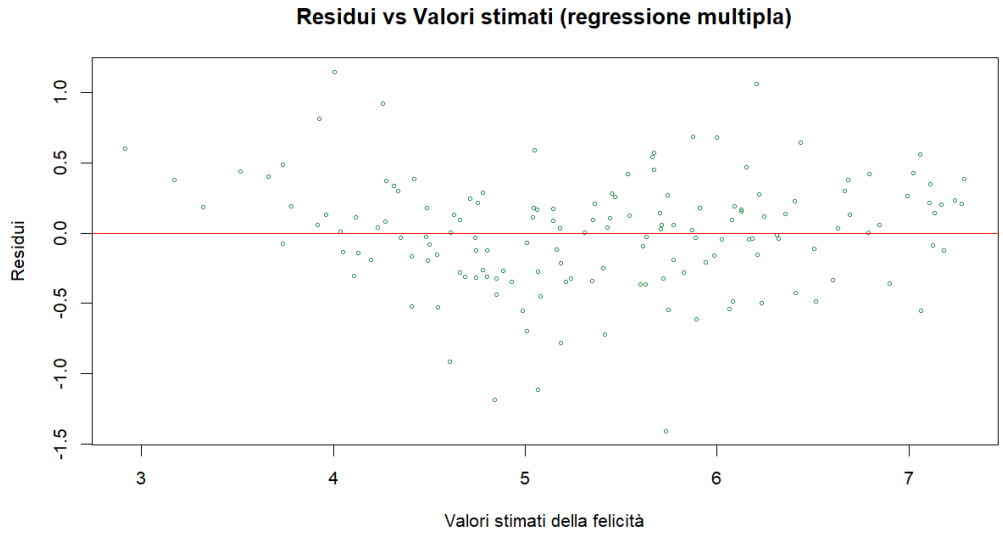


Figura 3.5: Residui nella regressione lineare multipla.

## 3.3 Clustering

Nelle analisi successive si cerca di identificare gruppi di paesi accomunati da caratteristiche socioeconomiche simili e valutare come, al variare dei fattori, i livelli di felicità cambino. Questa scelta è stata effettuata sulla base delle analisi precedenti, dove si è osservato come le variabili influenzino diversamente la percezione della felicità, considerando insiemi di Paesi con caratteristiche simili per determinare le variabili più impattanti. Tenendo conto delle osservazioni precedenti, si è deciso di escludere dal clustering la generosità, per evitare che appiattisca le differenze fra gruppi in quanto poco impattante sulla felicità e con basso potere discriminante.

### 3.3.1 Analisi delle componenti principali

Per individuare le variabili che più influenzano l'indice di felicità è stata utilizzata la PCA (Analisi delle componenti principali) tramite i valori standardizzati.

| | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** |
|---|---|---|---|---|---|---|---|
| Deviazione standard | 1.8370 | 1.1101 | 0.9376 | 0.8560 | 0.59718 | 0.52354 | 0.38773 |
| Varianza | 0.4821 | 0.1761 | 0.1256 | 0.1047 | 0.05095 | 0.03916 | 0.02148 |
| Proporzione cumulativa | 0.4821 | 0.6582 | 0.7837 | 0.8884 | 0.93937 | 0.97852 | 1.00000 |

Tabella 3.4: Importanza delle componenti principali

| **Variabile** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** |
|---|---|---|---|---|---|---|---|
| PIL pro capite | 0.440 | 0.450 | 0.021 | 0.063 | 0.056 | -0.120 | -0.763 |
| Supporto sociale | 0.444 | 0.170 | -0.365 | -0.157 | 0.102 | 0.746 | 0.224 |
| Emozioni positive | 0.337 | -0.478 | -0.028 | -0.536 | 0.540 | -0.277 | -0.049 |
| Emozioni negative | -0.278 | 0.325 | 0.685 | -0.462 | 0.195 | 0.310 | -0.023 |
| Aspettativa di vita sana | 0.414 | 0.483 | 0.151 | -0.047 | -0.053 | -0.464 | 0.594 |
| Libertà di scelta nella vita | 0.386 | -0.364 | 0.339 | -0.231 | -0.733 | 0.091 | -0.070 |
| Percezione della corruzione | -0.313 | 0.261 | -0.509 | -0.645 | -0.342 | -0.184 | -0.090 |

Tabella 3.5: Quanto ogni variabile standardizzata influisce sui componenti principali

Dall'analisi delle componenti principali risulta che:

- PC1 spiega il 48.21% della varianza totale ed è fortemente associata a variabili come PIL pro capite, Aspettativa di vita sana, Supporto sociale e Libertà di scelta nella vita. I livelli di corruzione e negatività sono invece bassi, rappresentando il benessere strutturale;
- PC2 contrappone variabili emotive e sociali (come Emozioni negative e positive) a quelle materiali (PIL pro capite e Aspettativa di vita sana), rappresentando come il benessere economico non coincide con una maggiore positività emotiva;
- PC3 è in gran parte influenzata da Emozioni negative ed esprime il malessere psicologico collettivo.

Le componenti successive spiegano quote minori di varianza ma rivelano:

- PC4 è influenzata dalla Percezione della corruzione con un contributo delle Emozioni, mostrando quindi la fiducia istituzionale;

- PC5 spiega la socialità culturale tramite Emozioni positive e Libertà di scelta nella vita;
- PC6 rappresenta l'equilibrio tra il Supporto sociale e l'Aspettativa di vita sana;
- PC7 è il bilanciamento fra Aspettativa di vita sana e Pil pro capite (salute e ricchezza);

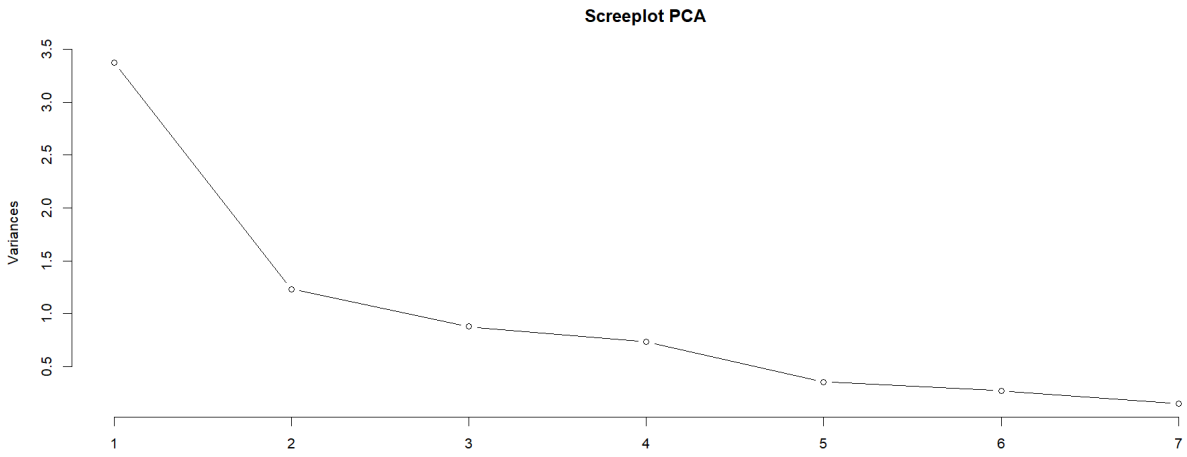


Figura 3.6: Screeplot PCA

Analizzando lo screeplot della PCA (Figura 3.6) si nota una forte differenza fra PC1 e PC2 confermando le osservazioni precedenti secondo le quali le prime osservazioni comprendono la maggior parte della varianza.

### 3.3.2 Divisione in cluster

Per l'analisi di similarità tra le variabili socioeconomiche è stato applicato un clustering gerarchico agglomerativo tramite il metodo del legame completo (Figura 3.7). Si fondono iterativamente le coppie di variabili più simili fino a ottenere un'unica struttura gerarchica (Tabella 3.6).

Per ottenere la matrice delle distanza su cui basare i cluster è stata utilizzata la metrica Euclidea sulle varaibili standardizzate.

$$d(\mathbf{X_i}, \mathbf{X_j}) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$$

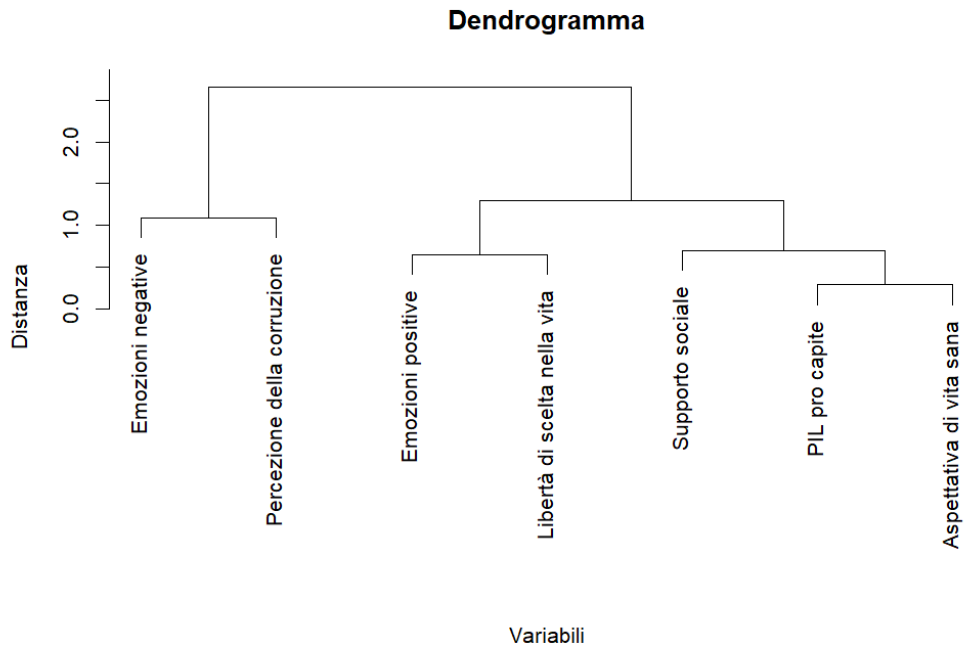


Figura 3.7: Dendrogramma delle variabili

Per semplicità nella tabella 3.6 è utilizzata la notazione:

- $i_1$ = PIL pro capite
- $i_2$ = Supporto sociale
- $i_3$ = Emozioni positive
- $i_4$ = Emozioni negative
- $i_5$ = Aspettativa di vita sana
- $i_6$ = Libertà di scelta nella vita
- $i_7$ = Percezione della corruzione

| N. cluster | Cluster | Livello di distanza |
|---|---|---|
| 7 | $\{i_1\}, \{i_2\}, \{i_3\}, \{i_4\}, \{i_5\}, \{i_6\}, \{i_7\}$ | 0.0000 |
| 6 | $\{i_1, i_5\}, \{i_2\}, \{i_3\}, \{i_4\}, \{i_6\}, \{i_7\}$ | 0.2852 |
| 5 | $\{i_1, i_5\}, \{i_2\}, \{i_3, i_6\}, \{i_4\}, \{i_7\}$ | 0.6461 |
| 4 | $\{i_1, i_2, i_5\}, \{i_3, i_6\}, \{i_4\}, \{i_7\}$ | 0.6984 |
| 3 | $\{i_1, i_2, i_5\}, \{i_3, i_6\}, \{i_4, i_7\}$ | 1.0931 |
| 2 | $\{i_1, i_2, i_3, i_5, i_6\}, \{i_4, i_7\}$ | 1.2963 |
| 1 | $\{i_1, i_2, i_3, i_4, i_5, i_6, i_7\}$ | 2.6534 |

Tabella 3.6: Sequenza di agglomerazioni del clustering gerarchico

Dall'analisi delle distanze si nota come esse aumentano progressivamente, evidenziando variabili fortemente correlate nelle prime aggregazioni, con un evidente salto fra 1 e 2 cluster e fra 2 e 3.

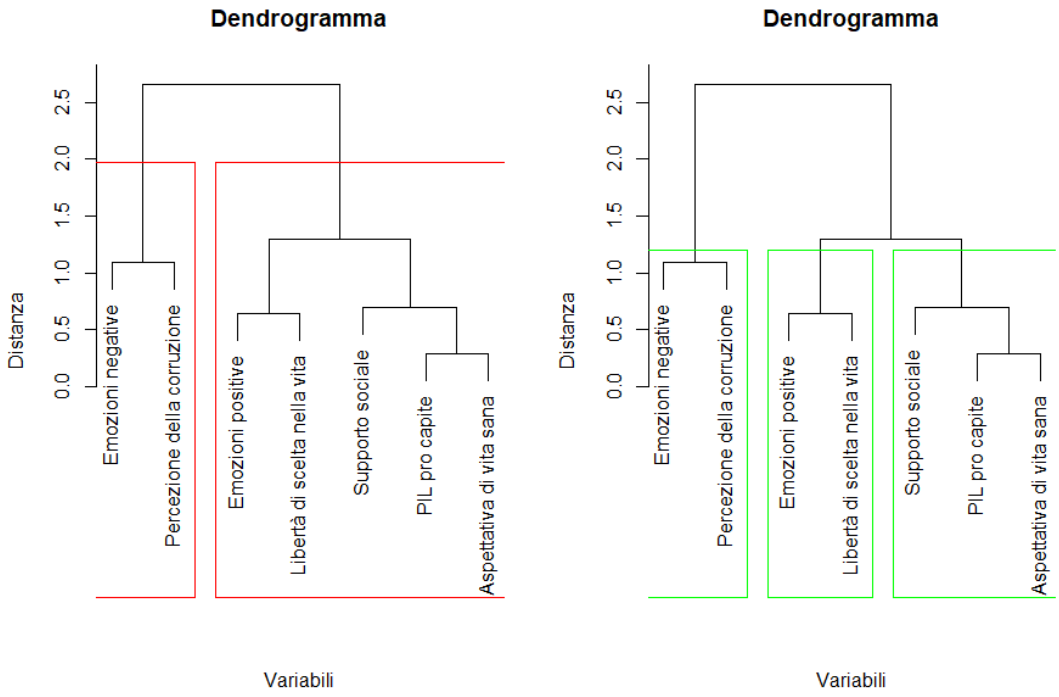


Figura 3.8: Dendrogramma della divisione fra clustering

Effettuando il taglio a 3 cluster si individuano tre macro aree:

- **Cluster 1 (benessere strutturale):** PIL pro capite, Supporto sociale e Aspettativa di vita sana rappresenta i fattori che spiegano la variazione della felicità.

- **Cluster 2 (dimensione psicologica positiva):** Emozioni positive e Libertà di scelta nella vita. Rappresenta aspetti positivi psicologici che influenzano la felicità.
- **Cluster 3 (malessere e sfiducia):** Percezione della corruzione e Emozioni negative. Rappresenta i fattori legati al disagio emotivo.

Effettuando il taglio a 2 cluster si individua una distinzione netta tra il benessere positivo e negativo:

- **Cluster 1 (benessere generale):** PIL pro capite, Aspettativa di vita sana, Supporto sociale, Emozioni positive e Libertà di scelta nella vita. Rappresenta i fattori che spiegano la variazione della felicità.
- **Cluster 2 (malessere):** Emozioni negative e Percezione della corruzione. Rappresenta i fattori legati al disagio emotivo.

Lo screeplot dei cluster trovati (Figura 3.9) mostra come la distanza maggiore è evidente nel passaggio fra 1 e 2 cluster, portando alle successive osservazioni per confermare la scelta del numero adatto di cluster (che forniscano gruppi con elementi appartenenti allo stesso gruppo più simili possibile e elementi di gruppi diversi più dissimili possibile).

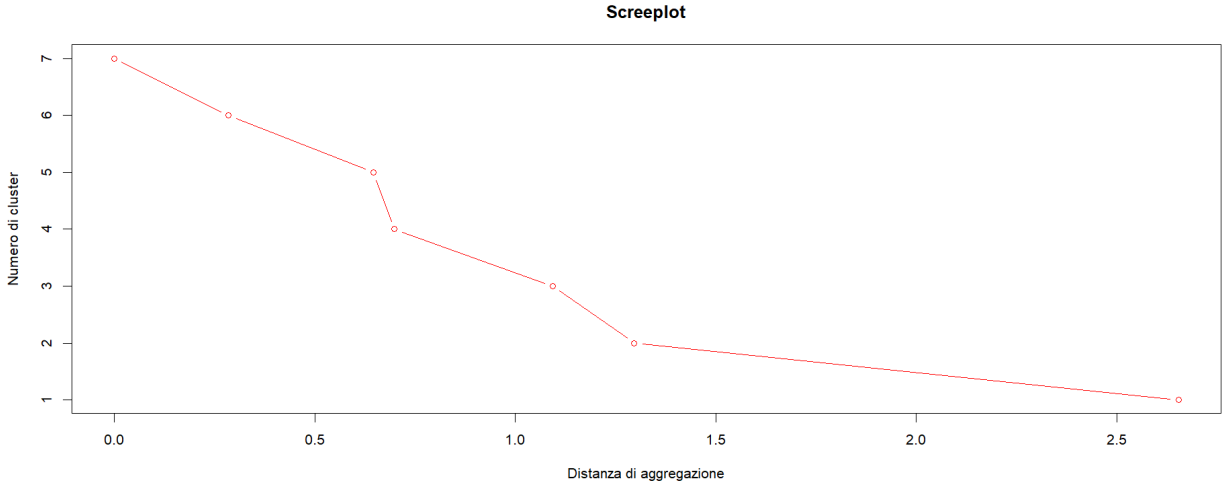


Figura 3.9: Screeplot dei cluster

Per scegliere il numero più adatto di cluster si è utilizzato il metodo della silhouette (Figura 3.10), evidenziando due cluster come divisione ottimale. (da qui le suddivisioni saranno sulla base di PC1 e PC3 che sarà rinominata PC2)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

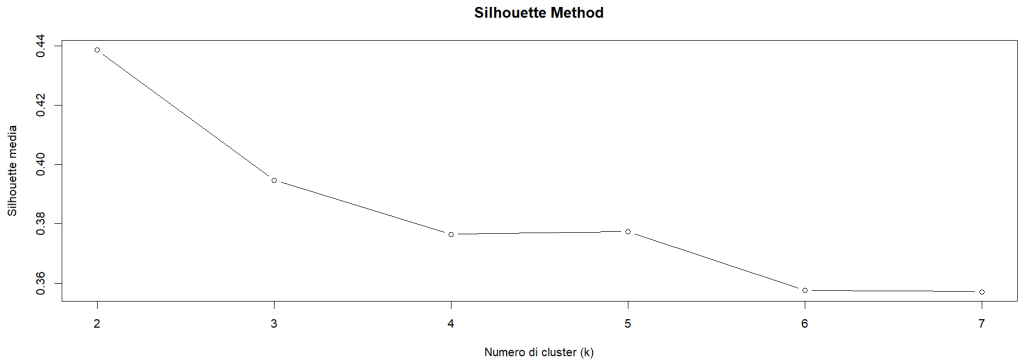


Figura 3.10: Metodo della silhouette

Si è poi proceduto con il clustering dei paesi con il metodo del legame completo (Figura 3.11) in due raggruppamenti.

$$D_{(i,j),k} = \max d(d_{ij}, d_{jk})$$

Per ogni paese è stato calcolato un profilo medio, viste le osservazioni che hanno portato alla conclusione che le variabili seguono un'andamento relativamente omogeneo per ogni paese nei vari anni (Figura 2.13).
I raggruppamenti ottenuti rappresentano quindi paesi caratterizzati da configurazioni simili nelle variabili importanti per ogni cluster.

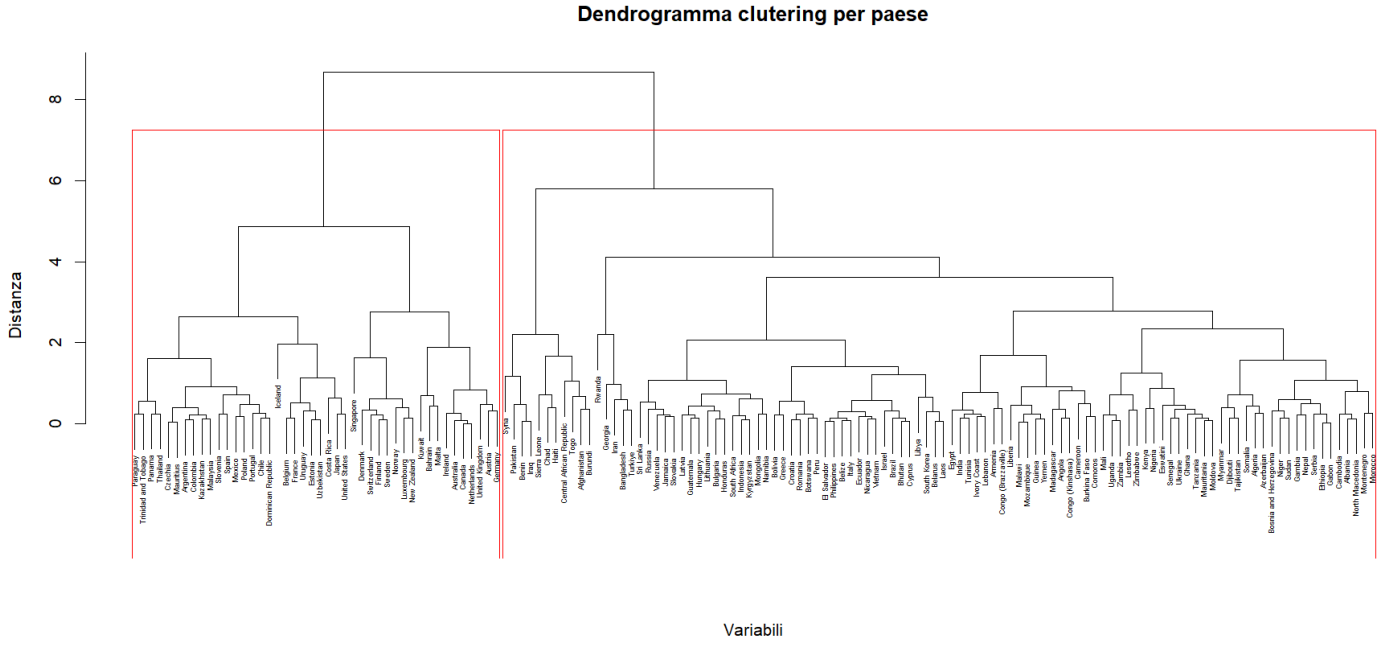


Figura 3.11: Dendrogramma del clustering per paese

### 3.3.3 Analisi della bontà dei Cluster

Dopo aver usato il clustering gerarchico per individuare il numero appropriato di cluster, è stato applicato il metodo del K-Means (con numero di cluster fissato a 2) per raggiungere una configurazione stabile fra cluster. Ottimizzando la geometria tra cluster essi vengono suddivisi in modo relativamente equilibrato (Figura 3.12), mostrando come la prima componente principale rappresenti la separazione più rilevante fra i gruppi.

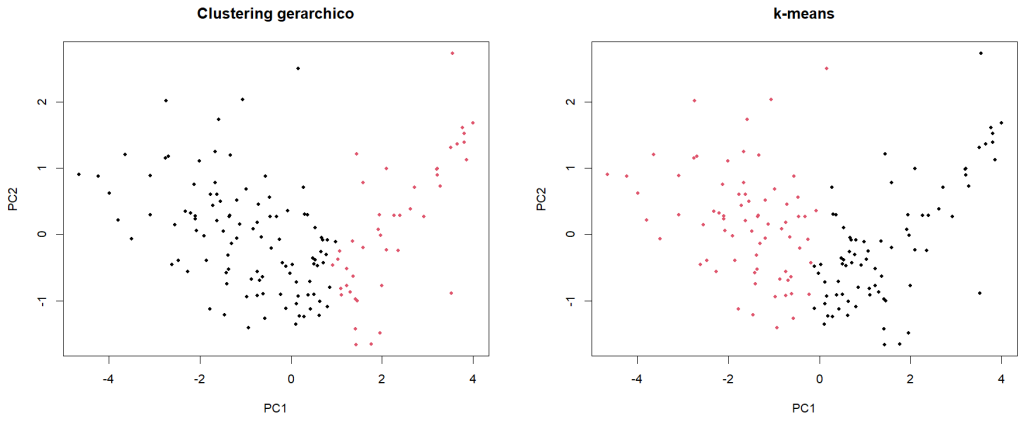


Figura 3.12: K-means

Per valutare la correttezza della divisione sono state analizzate varie metriche:

- **Within-Cluster Sum of Squares** (*292.6*) indica che i cluster sono dispersi con punti lontani dai propri centroidi;

$$WCSS = \sum_{j=1}^{K} \sum_{x \in C_k} d(x - \mu_j)^2$$

- **Between-Cluster Sum of Squares** (*354.8*) mostra una buona separazione fra i cluster, nonostante l'alto wcss;

$$BCSS = \sum_{j=1}^{K} |C_j| * d(\mu_j - \mu)^2$$

- **Calinski–Harabasz** (*177.04*) combina le metriche precedenti, indicando un buon bilanciamento fra i cluster.

$$CH = \frac{BCSS/(K-1)}{WCSS/(n-K)}$$

Tramite il metodo del gomito (Figura 3.13) è stata confermata la scelta di due cluster, in quanto l'aggiunta di ulteriori cluster non porterebbe a un miglioramento significativo della WCSS.

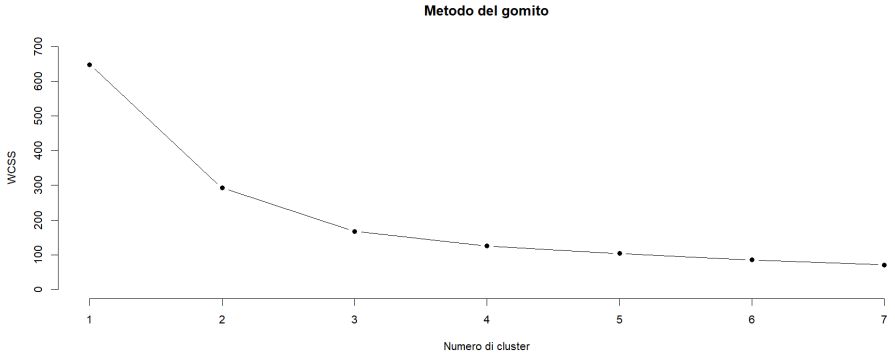


Figura 3.13: Metodo del gomito

I cluster evidenziano come la felicità sia influenzata sia da variabili positive che negative, permettendo di dividere il mondo in paesi felici ed infelici, come osservato analizzando il punteggio di felicità medio standardizzato fra i due cluster (Figura 3.14).

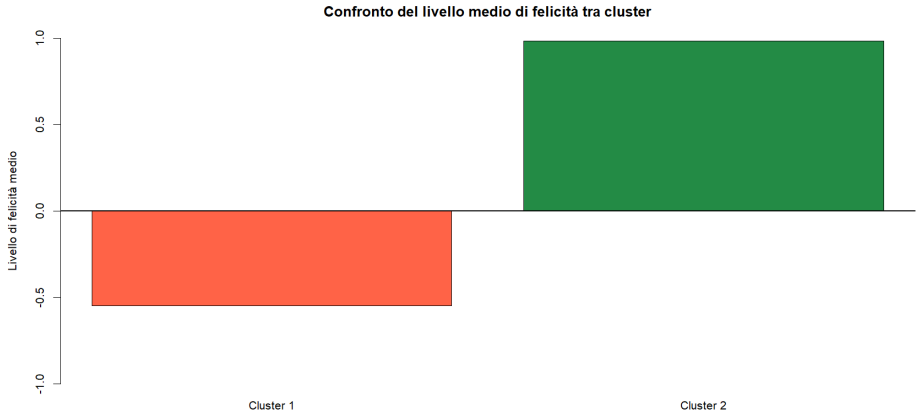


Figura 3.14: Analisi della media di felicità fra cluster

## 3.4 Analisi di inferenza

Le analisi esplorative e descrittive hanno permesso di individuare relazioni tra il punteggio di felicità e le variabili socioeconomiche. Per stabilire se le relazioni osservate rappresentino effetti statisticamente significativi nella popolazione dei paesi considerati e se ciò può essere applicato anche a pesi non presenti nelle rilevazioni, viene applicato un approccio di inferenza

statistica. Si testa quindi l'ipotesi di relazione sui parametri a partire dalle informazioni contenute nel campione, ponendo la felicità come variabile dipendente.

Per ciascuna variabile viene formulato un test di ipotesi sui coefficienti di regressione:

- $H_0$: L'ipotesi nulla indica l'assenza di effetto della variabile sulla felicità.
- $H_1$: L'ipotesi alternativa indica la presenza di una relazione statisticamente significativa.

Analizzando i compromessi si è scelto un livello di significatività pari a $\alpha = 0.05$ (test statisticamente significativo), cioè la probabilità di commettere un errore di tipo I, rifiutando l'ipotesi nulla quando essa è vera. La regione critica è stata determinata suddividendo $\alpha$ in due code simmetriche della distribuzione (Figura 3.15).

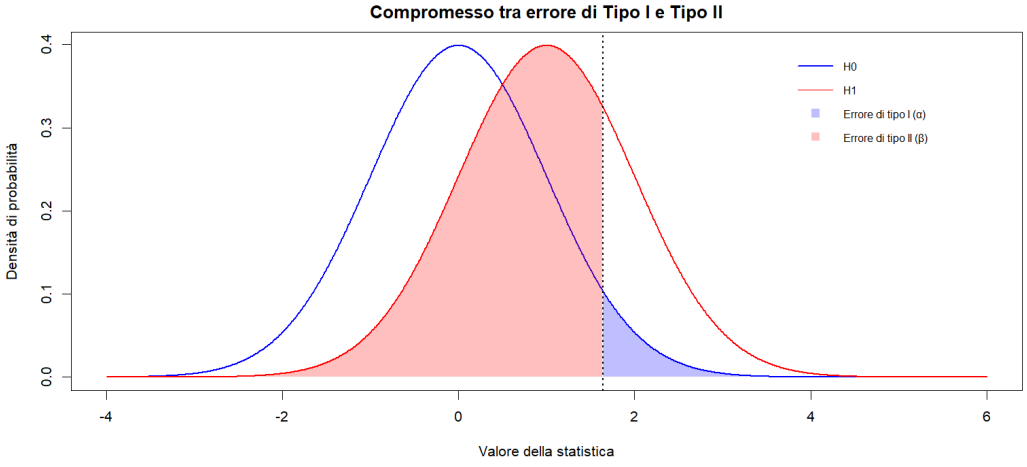


Figura 3.15: Alpha

| **Variabile** | $\hat{\beta}$ | Errore Std. | IC 95% | p-value | Segno |
|---|---|---|---|---|---|
| PIL pro capite | 0.754 | 0.041 | [0.673, 0.835] | $< 0.001$ | Positivo |
| Supporto sociale | 7.069 | 0.455 | [6.169, 7.969] | $< 0.001$ | Positivo |
| Aspettativa di vita sana | 0.117 | 0.007 | [0.103, 0.132] | $< 0.001$ | Positivo |
| Libertà di scelta nella vita | 5.442 | 0.573 | [4.310, 6.574] | $< 0.001$ | Positivo |
| Generosità | 1.281 | 0.605 | [0.086, 2.476] | 0.036 | Positivo |
| Percezione della corruzione | -2.785 | 0.463 | [-3.701, -1.869] | $< 0.001$ | Negativo |
| Emozioni positive | 5.880 | 0.772 | [4.354, 7.406] | $< 0.001$ | Positivo |
| Emozioni negative | -6.055 | 1.181 | [-8.389, -3.721] | $< 0.001$ | Negativo |

Tabella 3.7: Risultati delle regressioni lineari semplici tra felicità e variabili

Per valutare l'influenza delle singole variabili socioeconomiche sul punteggio medio di felicità, è stato svolto un test bilaterale ciascuna variabile (Tabella 3.7). I risultati evidenziano che tutte le variabili considerate presentano un coefficiente statisticamente significativo, con intervalli di confidenza che non includono il valore zero.
Il coefficiente di regressione mostra di quanto cambia la felicità aumentando di 1 unità la variabile, evidenziando come le variabili meno influenti siano il Pil pro capite, l'aspettativa di vita sana e la generosità (nonostante questo coefficiente non sia l'unico fattore da considerare).
L'errore standard relativamente piccolo mostra una stima abbastanza affidabile, con un valore anomalo nelle emozioni negative. Il p-value inferiore al livello di significatività permette quindi di rifiutare l'ipotesi nulla per tutte le variabili. Solamente la generosità si mostra meno correlata e impattante, avendo un valore meno significativo rispetto agli altri.

L'analisi inferenziale conferma quindi l'esistenza di relazioni tra le variabili socioeconomiche analizzate e il livello di felicità, come evidenziato dalle analisi precedenti.

## 3.5 Conclusione

Dopo aver svolto le opportune analisi si è osservato che la felicità non è influenzata solo da fattori economici, ma anche sociali e psicologici.

L'analisi inferenziale conferma che tutte le variabili, a eccezione della generosità, risultano essere fortemente e mediamente associate alla felicità media dei paesi. Le variabili legate alla percezione della corruzione e agli affetti negativi presentano invece un effetto negativo significativo.
I paesi possono quindi essere divisi in felici e infelici a seconda dei principali gruppi di variabili (positive e negative), mostrando la loro influenza sul benessere emotivo di una popolazione.

# Research question 2

***I dati generati da LLM (in particolare ChatGPT e Gemini) possono essere usati per migliorare la qualità del dataset reale? (es. integrando i dati per gli anni in cui non ci sono rilevazioni per determinati paesi)***
***In particolare, quale livello di coerenza statistica e strutturale tali dati raggiungono rispetto al dataset reale, considerando la distribuzione delle feature, la valutazione delle principali tecniche di imputazione e i potenziali rischi di data leak.***

## 4.1 Generazione di un dataset

Per entrambi i modelli, ovvero ChatGPT e Gemini, è stato utilizzato un prompt strutturato per generare i dati in modo più simile possibile al dataset reale e ai dati mondiali. Al modello viene richiesto di:

- Generare un dataset sintetico realistico basandosi su elementi conosciuti senza cercare informazioni esterne;
- Per evitare data leak è vietato l'utilizzo di World Happiness Report (inclusi pattern, ranking, valori medi o distribuzioni specifiche). I dati devono essere plausibili ma non ricostruibili a partire dal World Happiness Report;
- Utilizzare solo i nomi delle variabili e i paesi del dataset originale e generare dati nel range di anni dal 2005 al 2022;
- Ogni paese deve mantenere coerenza temporale interna (trend realistici, assenza di salti implausibili);
- Generare l'output in un file csv.

### 4.1.1 ChatGPT

L'analisi grafica del dataset generato tramite ChatGPT evidenzia pattern regolari.

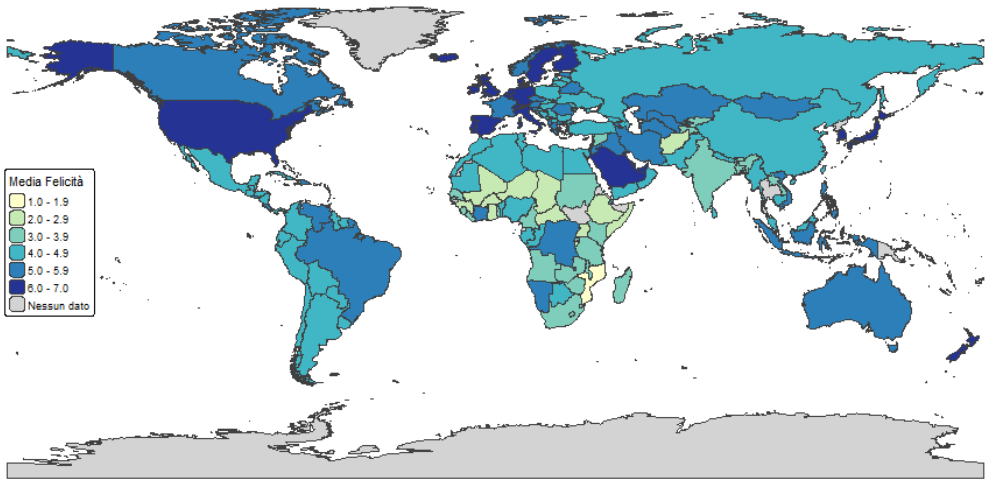


Figura 4.1: Mappa della felicità mondiale nel dataset generato da ChatGPT.

Osservando la mappa della felicità dei valori generati da ChatGPT (Figura 4.1) si nota che i valori sono più elevati rispetto a quelli originali, mostrando una netta differenza.

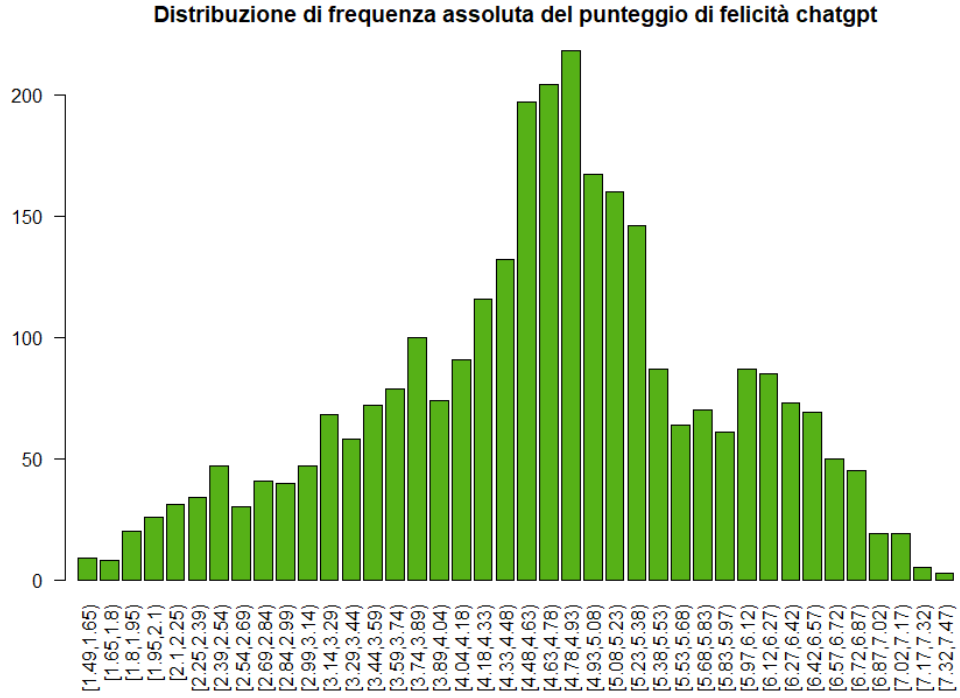


Figura 4.2: Distribuzione della frequenza assoluta di felicità generato da ChatGPT.

Osservando la distribuzione della frequenza assoluta dei valori generati da ChatGPT (Figura 4.2) la distribuzione risulta unimodale, quasi simmetrica con picco tra 4.5 e 5.0. Si

presenta complessivamente bilanciata, aumentando il numero di valori bassi rispetto a quelli alti originariamente presenti.

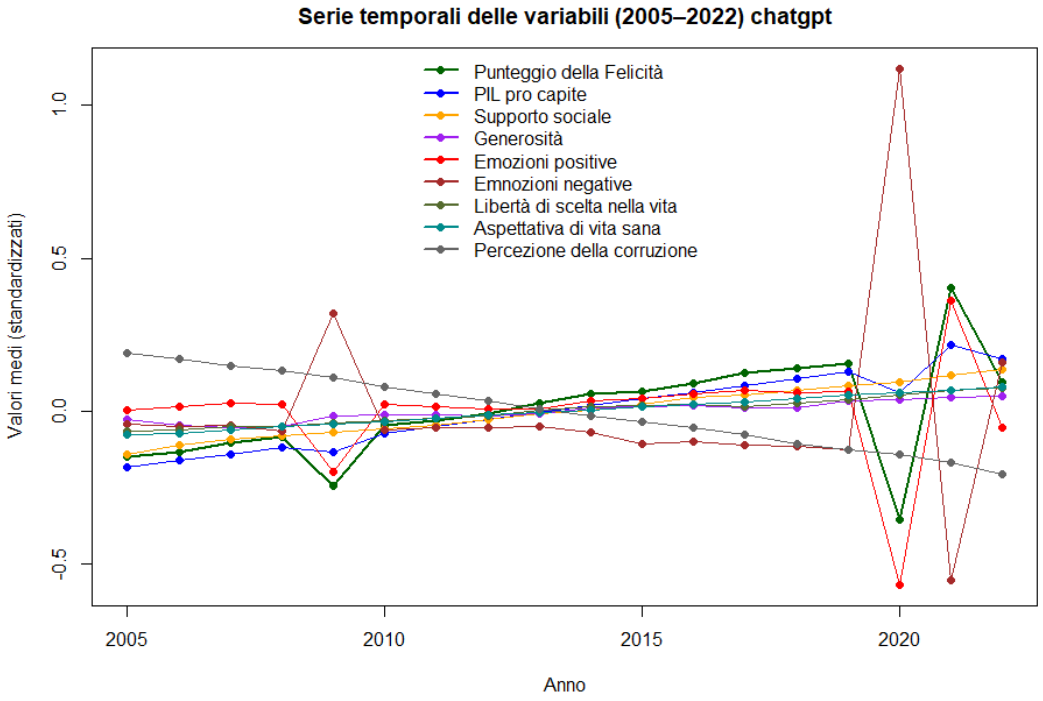


Figura 4.3: Serie temporare multivariata generato da ChatGPT della felicità.

Osservando la serie temporale (Figura 4.3) si nota come i dati presentano pattern simmetrici con centro di simmetria gli anni 2013-2014. La distribuzione è quindi prevedibile e le variabili risultano artificiali e molto diverse da una distribuzione naturale. Alcuni punti non sono prevedibili, come l'anno 2009 e dopo il 2020. Ciò potrebbe essere influenzato da una conoscenza del modello di fenomeni globali, come l'epidemia del 2020 e la conseguente crisi economica.

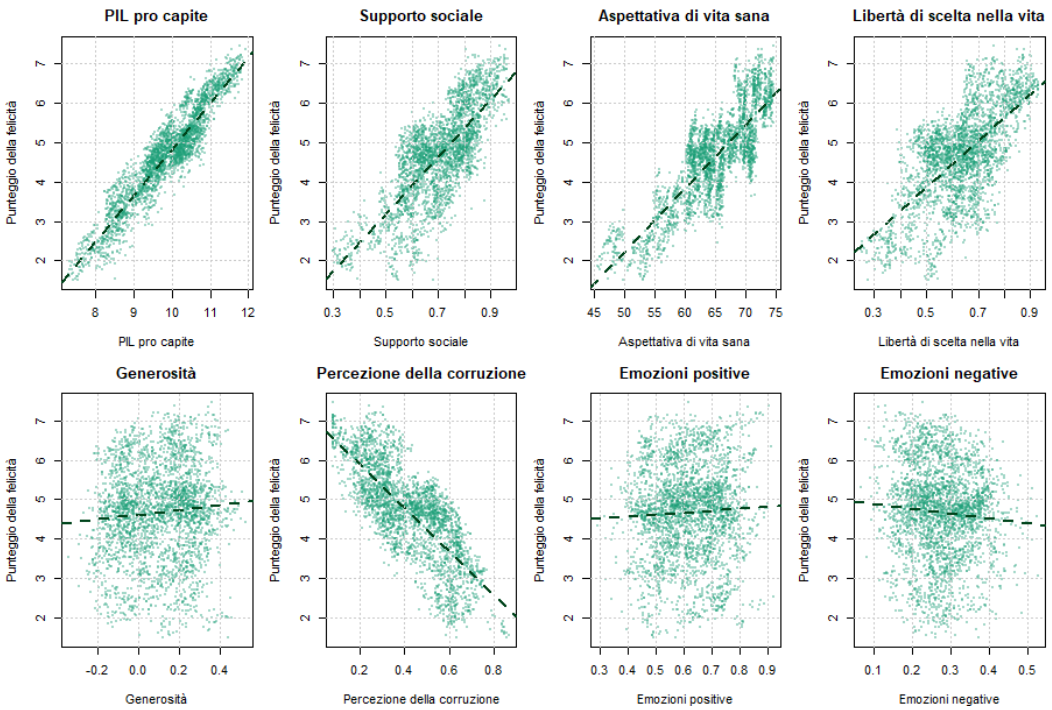


Figura 4.4: Scatterplot delle relazioni delle variabili con la felicità.

Analizzando la relazione fra le variabili e l'indice di felicità (Figura 4.4) si notano pattern quasi lineari. Il dataset sintetico presenta addensamenti evidenti e, per alcune variabili, non c'è più la relazione che si osservava invece fra le variabili originali e la felicità.

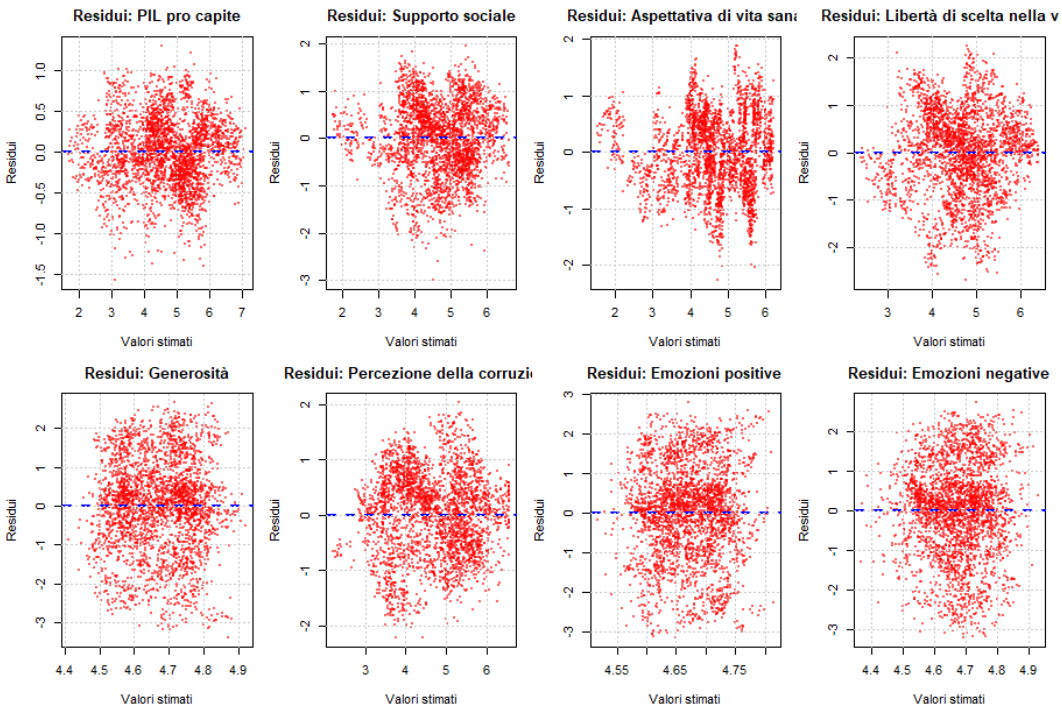


Figura 4.5: Scatterplot dei residui.

L'analisi dei residui (Figura 4.5) mostra come i dati generati da ChatGPT non raggiungono un livello di coerenza statistica e strutturale comparabile a quello del dataset reale. Seppure in alcuni punti è visibile una distribuzione dei residui quasi reale, altri presentano un pattern visibile. L'uso di tali dati introduce distorsioni nelle misure di correlazione e nei modelli di regressione, compromettendo l'interpretabilità dei risultati.

### 4.1.2 Gemini

L'analisi grafica del dataset generato tramite Gemini presenta risultati meno realistici rispetto a quelli di ChatGPT e comunque con pattern regolari.

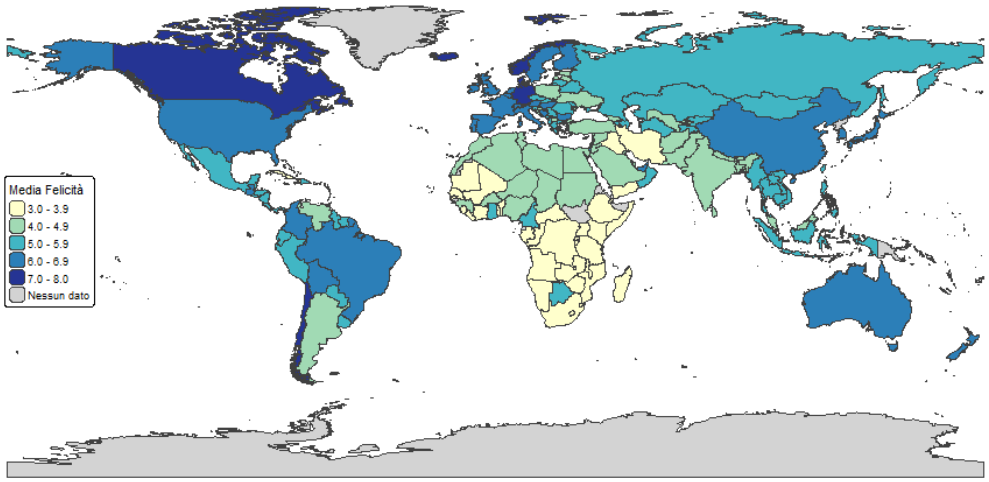


Figura 4.6: Mappa della felicità mondiale nel dataset generato da Gemini.

Osservando la mappa della felicità dei valori generati da Gemini (Figura 4.6) si nota che i paesi africani e dell'ovest tendono ad avere una felicità media più bassa rispetto all'originale e quasi nessun paese ottiene un incremento, come confermato dalle analisi successive.

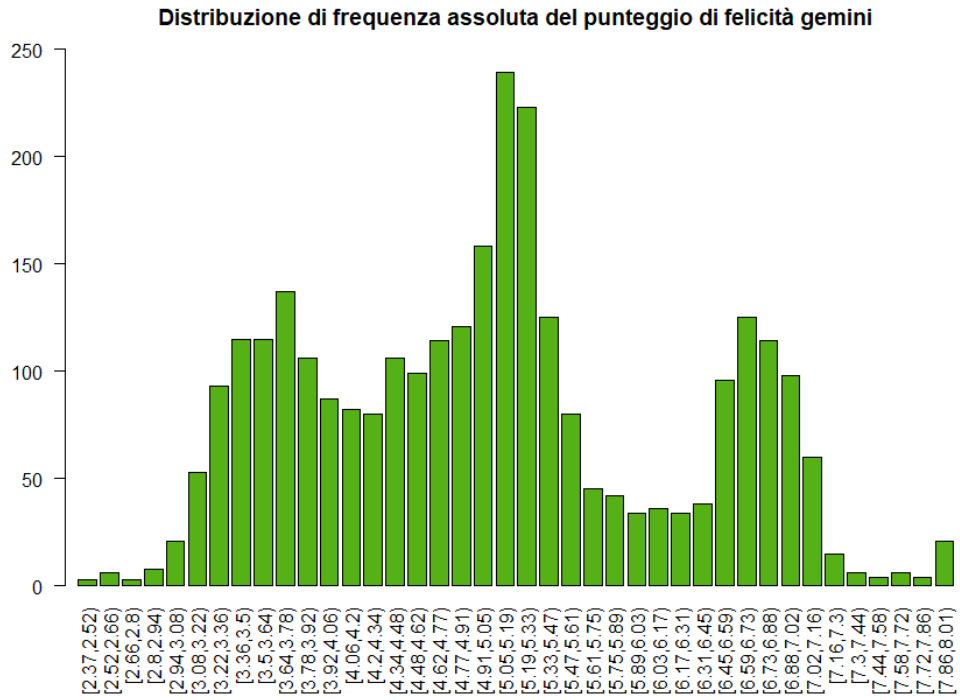


Figura 4.7: Distribuzione della frequenza assoluta di felicità generato da Gemini.

Osservando la distribuzione della frequenza assoluta (Figura 4.7) la distribuzione risulta irregolare e multimodale con 3 picchi: da 3.5 a 4.0, il più elevato attorno a 5.0 e l'ultimo

tra 6.5 a 7.0. La distribuzione risulta meno realistica, in quanto dai valori medi si passa a valori più elevati ma pochi paesi hanno un livello di felicità superiore al 7. I valori inferiori, invece, sono in maggioranza.

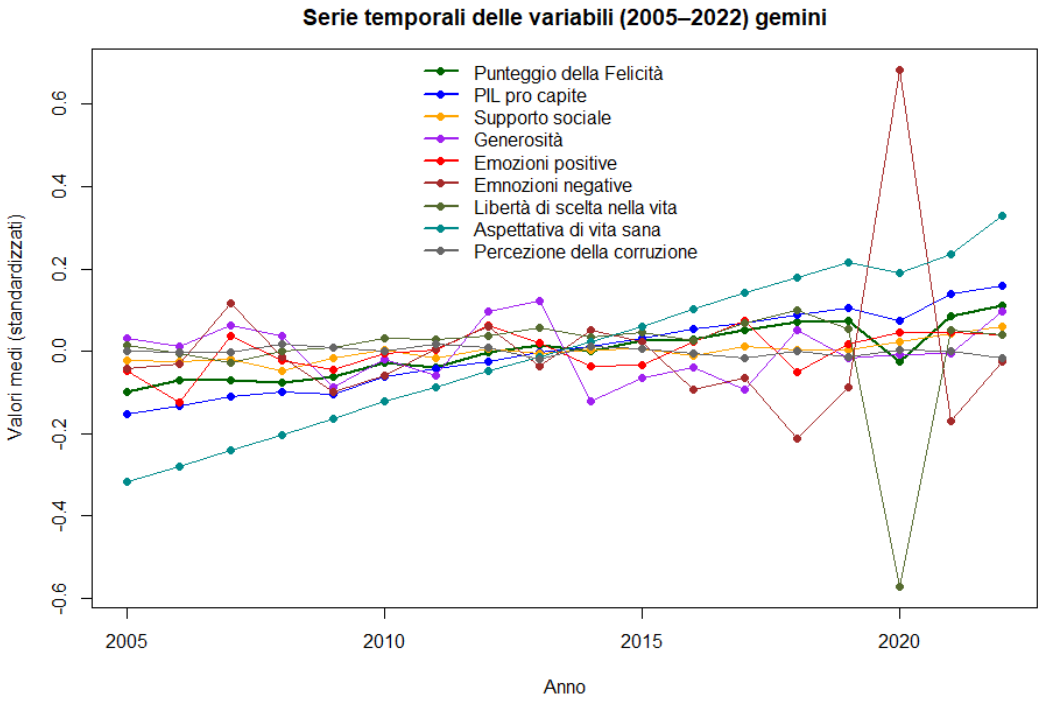


Figura 4.8: Serie temporale multivariata generato da Gemini della felicità.

La serie temporale generata da Gemini (Figura 4.8) presenta evidenti anomalie. Ciò potrebbe essere dovuto all'influenza di fonti esterne che, mostrando al modello un periodo di forte crisi nell'anno 2020, hanno portato a una variazione estrema dei dati.

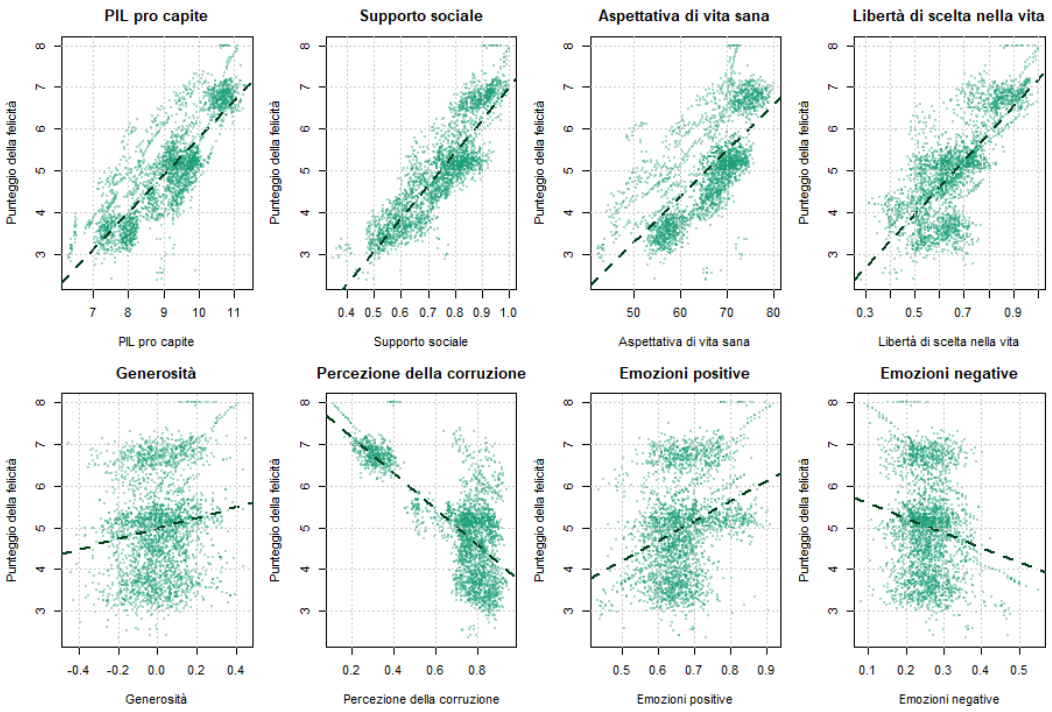


Figura 4.9: Scatterplot delle relazioni delle variabili con la felicità.

Dall'osservazione degli scatterplot (Figura 4.9), emerge che Gemini non riesce a fornire realismo come ChatGPT. Sono presenti criticità come Emozioni positive, Emozioni negative e Generosità che presentano pattern regolari. La correlazione delle variabili rispetto alla felicità è per la maggior parte elevata, sia in ambito positivo che negativo.

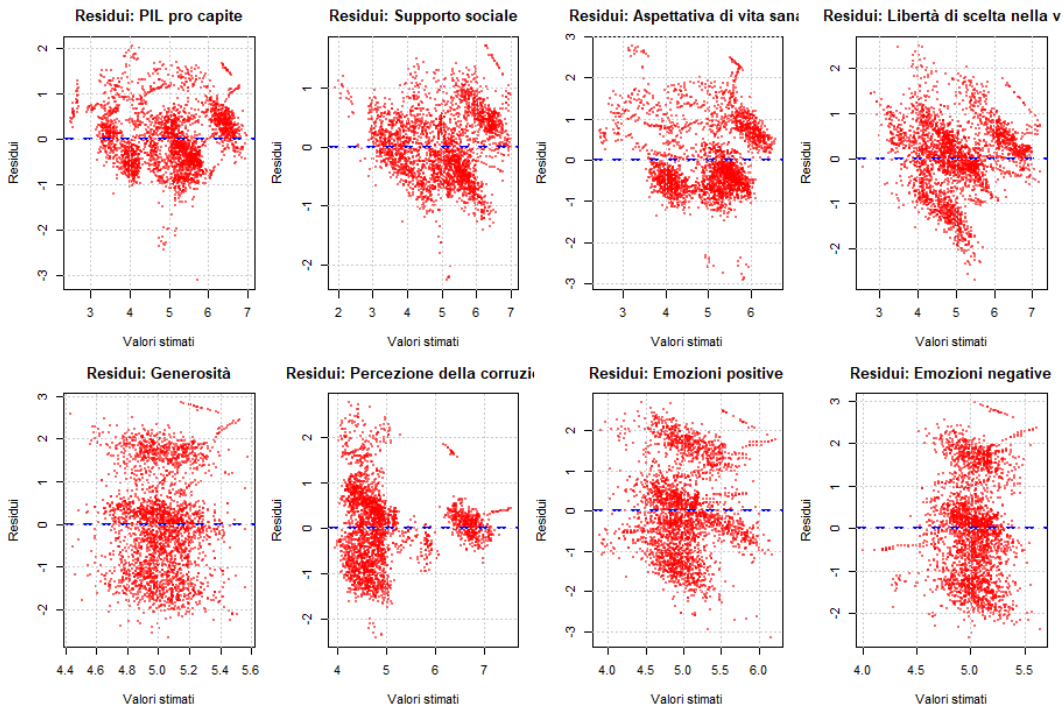


Figura 4.10: Scatterplot dei residui.

Il diagramma dei residui (Figura 4.10) conferma le problematiche riscontrate precedentemente, ovvero i residui non sono distribuiti in modo completamente casuale ma sono organizzati in regioni definite nello spazio.

Gemini, a differenza di ChatGPT, presenta pattern artificiali marcati con una predisposizione verso valori negativi.

## 4.2 Completamento del dataset originale

Osservando il dataset originale (Tabella 2.1) sono stati riscontrati valori mancanti. Si è quindi cercato di ricostruirli tramite l'utilizzo di ChatGPT, cercando di comprendere se l'uso di dati sintetici sia un valido sostituto a valori ricostruiti statisticamente.

| Country | Code | Year | Happiness | Log GDP | Social | Health | Freedom | Generosity | Corruption | Pos. Affect | Neg. Affect |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | | 2005 | 3.724 | 7.35 | 0.451 | 50.5 | 0.718 | 0.168 | 0.882 | 0.414 | 0.258 |
| Afghanistan | | 2006 | 3.724 | 7.35 | 0.451 | 50.5 | 0.718 | 0.168 | 0.882 | 0.414 | 0.258 |
| Afghanistan | | 2007 | 3.724 | 7.35 | 0.451 | 50.5 | 0.718 | 0.168 | 0.882 | 0.414 | 0.258 |
| Afghanistan | AFG | 2008 | 3.724 | 7.35 | 0.451 | 50.5 | 0.718 | 0.168 | 0.882 | 0.414 | 0.258 |
| Afghanistan | AFG | 2009 | 4.402 | 7.509 | 0.552 | 50.8 | 0.679 | 0.191 | 0.850 | 0.481 | 0.237 |

Tabella 4.1: Estratto del dataset modificato

| Country | Code | Year | Happiness | Log GDP | Social | Health | Freedom | Generosity | Corruption | Pos. Affect | Neg. Affect |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Armenia | ARM | 2017 | 4.288 | 9.434 | 0.698 | 66.55 | 0.614 | -0.152 | 0.865 | 0.552 | 0.437 |
| Armenia | ARM | 2018 | 5.062 | 9.490 | 0.814 | 66.83 | 0.808 | -0.169 | 0.677 | 0.535 | 0.455 |
| Armenia | ARM | 2019 | 5.488 | 9.569 | 0.782 | 67.10 | 0.844 | -0.179 | 0.583 | 0.537 | 0.430 |
| Armenia | | 2020 | 5.395 | 9.565 | 0.772 | 67.38 | 0.820 | -0.167 | 0.644 | 0.552 | 0.454 |
| Armenia | ARM | 2021 | 5.301 | 9.561 | 0.762 | 67.65 | 0.795 | -0.156 | 0.705 | 0.566 | 0.478 |

Tabella 4.2: Estratto del dataset modificato

Osservando i dati ricostruiti da ChatGPT si evidenziano degli errori:

- Non è presente alcun dato nella sezione Code, quindi il codice di avviamento postale non viene generato;
- Alcune righe ricostruite sono copie di una riga del paese presente nel dataset originale (Tabella 4.1);
- In alcuni casi la riga è generata con dati graduali (Tabella 4.2).

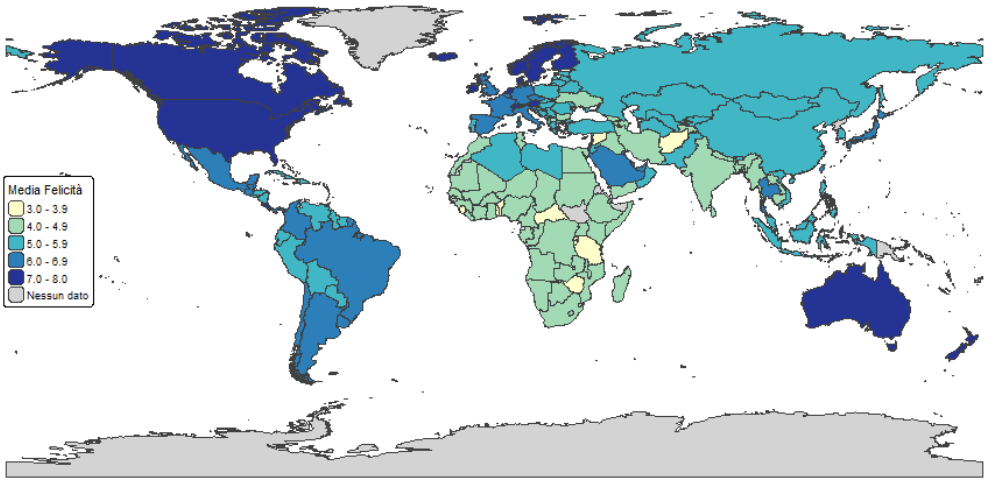


Figura 4.11: Mappa della felicità mondiale nel dataset originale completato da ChatGPT.

Comparando la mappa originale (Figura 2.1) con quella creata tramite il dataset completato artificialmente (Figura 4.11) si nota che le differenze della media della felicità sono minime, tranne che nei paesi africani dove la differenza è evidente.

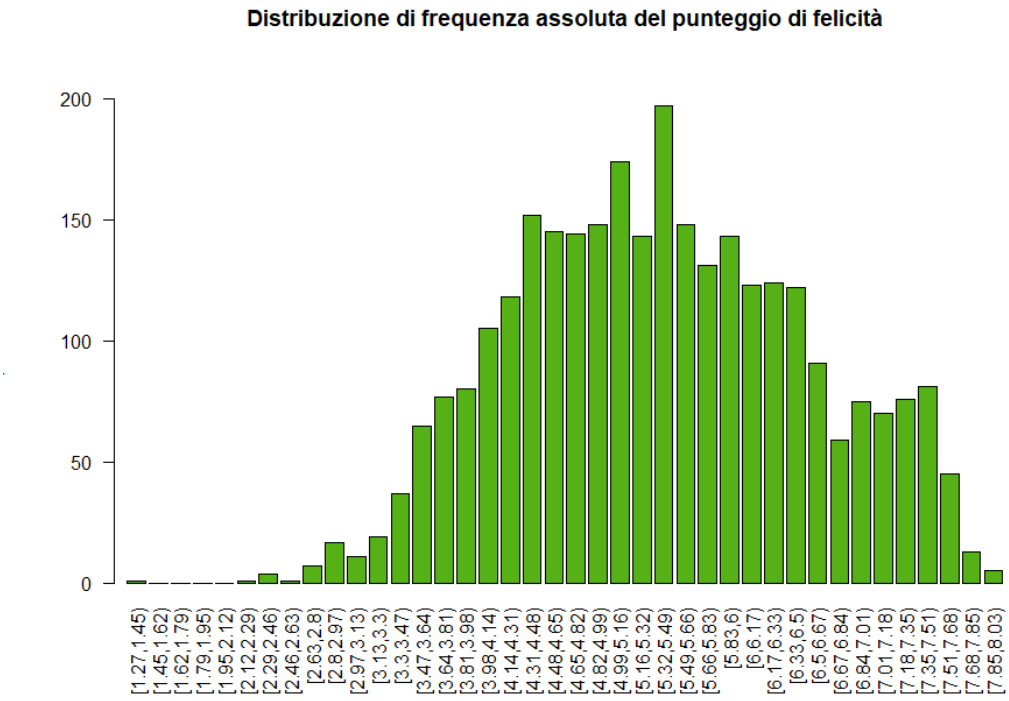


Figura 4.12: Distribuzione della frequenza assoluta di felicità del dataset originale completato da ChatGPT.

Osservando la distribuzione della frequenza assoluta (Figura 4.12) la distribuzione è quasi identica all'originale (Figura 2.5), con tuttavia un appiattimento dei valori inferiori, essendo unimodale con picco tra 4.8 e 5.3.

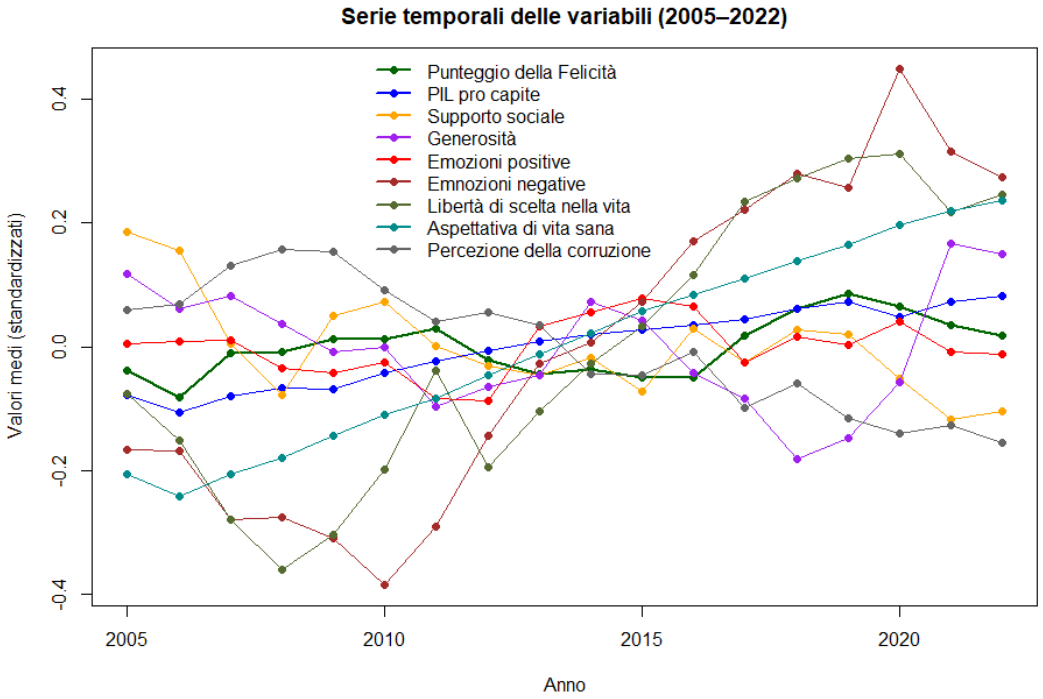


Figura 4.13: Serie temporale multivariata del dataset originale completato da ChatGPT della felicità.

Confrontando la serie temporale del dataset iniziale (Figura 2.13) con quella creata tramite il dataset completato artificialmente (Figura 4.13)) i valori anomali dell'anno 2005 presenti nel dataset originale non sono più presenti e i nuovi risultano ben distribuiti. Ad essere evidente è Emozioni Negative dove dapprima presenta una scesa, ma dal 2010 è presente una forte risalita con picco l'anno 2020 (probabilmente fattore scatenante è il COVID-19). Un altro valore diversificato è Aspettativa di vita sana che risulta salire gradualmente nella serie, senza alcuna alterazione.

## 4.3 Conclusione

Entrambi gli LLM utilizzati (ChatGPT e Gemini) non sono adatti a generare un dataset fedele a dati originali basandosi su conoscenze apprese. Le relazioni tra le variabili risultano lineari e deterministiche, con evidenti pattern.

I dati generati, quindi, non possono essere considerati un valido sostituto o complemento diretto dei dati reali, ma al più uno strumento esplorativo.

Utilizzando ChatGPT per completare i dati mancanti del dataset esistente si è notata un'alterazione dei dati originali, oltre all'aggiunta dei dati richiesti. Ciò ha portato alla modifica dei valori del dataset e a risultati delle analisi successive leggermente diversi rispetto

all'originale. Per questo utilizzare un modello esistente può non essere la scelta adeguata per completare i dati di un dataset rispetto a, invece, soluzioni di imputazione statistica come quelle utilizzate.

# Note

Le analisi sono state svolte in R e sono consultabili al seguente link:
`https://github.com/YuliaD2609/World_happiness_report.git`