DSLAB

2025-07-01

```
library(readxl)
```

```
## Warning: il pacchetto 'readxl' è stato creato con R versione 4.2.3
```

```
percorso_file_principale <- "D:/Dropbox/Desktop/Università/Magistrale/dataset1DSLAB.xlsx"

dati_score_topic <- read_excel(percorso_file_principale, sheet = "Score_Last", skip = 5)

dati_score_topic <- dati_score_topic[-c(1, 449:496), ]

head(dati_score_topic)
```

```
## # A tibble: 6 x 14
##   Country   Region        Code  Education Jobs  Income Safety Health Environment
##   <chr>     <chr>         <chr> <chr>     <chr> <chr>  <chr>  <chr>  <chr>
## 1 Australia New South Wa~ AU1   8.507677~ 8.23~ 5.683~ 9.809~ 9.470~ 7.56218905~
## 2 Australia Victoria      AU2   8.507677~ 8.42~ 4.604~ 9.755~ 9.945~ 8.30845771~
## 3 Australia Queensland    AU3   8.192631~ 8.34~ 5.017~ 9.782~ 9.212~ 9.30348258~
## 4 Australia South Austra~ AU4   7.579120~ 8.09~ 5.303~ 9.918~ 9.097~ 9.60199004~
## 5 Australia Western Aust~ AU5   8.192631~ 8.85~ 5.910~ 9.755~ 9.619~ 8.45771144~
## 6 Australia Tasmania      AU6   7.230912~ 8.06~ 4.938~ 9.864~ 8.257~ 9.00497512~
## # i 5 more variables: `Civic engagement` <chr>,
## #   `Accessiblity to services` <chr>, Housing <chr>, Community <chr>,
## #   `Life satisfaction` <chr>
```

```
require(skimr)
```

```
## Caricamento del pacchetto richiesto: skimr
```

```
## Warning: il pacchetto 'skimr' è stato creato con R versione 4.2.3
```

```
skim_without_charts(dati_score_topic)
```

Table 1: Data summary

Name	dati_score_topic
Number of rows	447
Number of columns	14
Column type frequency:	

character	14
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Country	0	1	5	15	0	38	0
Region	0	1	4	35	0	447	0
Code	0	1	3	5	0	447	0
Education	0	1	2	21	0	351	0
Jobs	0	1	2	21	0	426	0
Income	0	1	2	21	0	433	0
Safety	0	1	2	21	0	139	0
Health	0	1	1	21	0	443	0
Environment	0	1	2	21	0	177	0
Civic engagement	0	1	2	21	0	401	0
Accessiblity to services	0	1	14	21	0	447	0
Housing	0	1	1	21	0	160	0
Community	0	1	1	21	0	192	0
Life satisfaction	0	1	1	21	0	42	0

```
str(dati_score_topic)
```

```
## tibble [447 x 14] (S3: tbl_df/tbl/data.frame)
##  $ Country                 : chr [1:447] "Australia" "Australia" "Australia" "Australia" ...
##  $ Region                  : chr [1:447] "New South Wales" "Victoria" "Queensland" "South Australia"
##  $ Code                    : chr [1:447] "AU1" "AU2" "AU3" "AU4" ...
##  $ Education               : chr [1:447] "8.5076771281666996" "8.5076771281666996" "8.192631188557449
##  $ Jobs                    : chr [1:447] "8.2346132128740805" "8.4209486166007892" "8.346320346320350
##  $ Income                  : chr [1:447] "5.6833387728459499" "4.6042754569190603" "5.01713446475196
##  $ Safety                  : chr [1:447] "9.8097826086956506" "9.7554347826087007" "9.782608695652170
##  $ Health                  : chr [1:447] "9.4703290246768503" "9.94566136508951" "9.21269095182139"
##  $ Environment             : chr [1:447] "7.5621890547263702" "8.3084577114427898" "9.303482587064680
##  $ Civic engagement        : chr [1:447] "10.0098914930556" "10.009892031958501" "10.0098903749178"
##  $ Accessiblity to services: chr [1:447] "5.37556933715112" "5.5247313046368998" "5.2789859130123897
##  $ Housing                 : chr [1:447] "7.9213483146067398" "9.6067415730337107" "7.921348314606739
##  $ Community               : chr [1:447] "8.8738738738738707" "8.8738738738738707" "7.072072072072069
##  $ Life satisfaction       : chr [1:447] "8.4615384615384599" "8.8461538461538503" "8.076923076923080
```

```
dati_score_topic$Education <- as.numeric(dati_score_topic$Education)
```

```
## Warning: NA introdotti per coercizione
```

```
dati_score_topic$Jobs <- as.numeric(dati_score_topic$Jobs)
```

```
## Warning: NA introdotti per coercizione
```

```
dati_score_topic$Income <- as.numeric(dati_score_topic$Income)
```

```
## Warning: NA introdotti per coercizione
```

```
dati_score_topic$Safety <- as.numeric(dati_score_topic$Safety)
```

```
## Warning: NA introdotti per coercizione
```

```
dati_score_topic$Health <- as.numeric(dati_score_topic$Health)
dati_score_topic$Environment <- as.numeric(dati_score_topic$Environment)
dati_score_topic$`Civic engagement` <- as.numeric(dati_score_topic$`Civic engagement`)
```

```
## Warning: NA introdotti per coercizione
```

```
dati_score_topic$`Accessiblity to services` <- as.numeric(dati_score_topic$`Accessiblity to services`)
dati_score_topic$Housing <- as.numeric(dati_score_topic$Housing)
```

```
## Warning: NA introdotti per coercizione
```

```
dati_score_topic$Community <- as.numeric(dati_score_topic$Community)
```

```
## Warning: NA introdotti per coercizione
```

```
dati_score_topic$`Life satisfaction` <- as.numeric(dati_score_topic$`Life satisfaction`)
```

```
## Warning: NA introdotti per coercizione
```

```
str(dati_score_topic)
```

```
## tibble [447 x 14] (S3: tbl_df/tbl/data.frame)
##  $ Country                 : chr [1:447] "Australia" "Australia" "Australia" "Australia" ...
##  $ Region                  : chr [1:447] "New South Wales" "Victoria" "Queensland" "South Australia"
##  $ Code                    : chr [1:447] "AU1" "AU2" "AU3" "AU4" ...
##  $ Education               : num [1:447] 8.51 8.51 8.19 7.58 8.19 ...
##  $ Jobs                    : num [1:447] 8.23 8.42 8.35 8.1 8.85 ...
##  $ Income                  : num [1:447] 5.68 4.6 5.02 5.3 5.91 ...
##  $ Safety                  : num [1:447] 9.81 9.76 9.78 9.92 9.76 ...
##  $ Health                  : num [1:447] 9.47 9.95 9.21 9.1 9.62 ...
##  $ Environment             : num [1:447] 7.56 8.31 9.3 9.6 8.46 ...
##  $ Civic engagement        : num [1:447] 10 10 10 10 10 ...
##  $ Accessiblity to services: num [1:447] 5.38 5.52 5.28 4.87 5.16 ...
##  $ Housing                 : num [1:447] 7.92 9.61 7.92 7.92 9.61 ...
##  $ Community               : num [1:447] 8.87 8.87 7.07 8.38 7.61 ...
##  $ Life satisfaction       : num [1:447] 8.46 8.85 8.08 8.46 8.46 ...
```

```
require(skimr)
skim_without_charts(dati_score_topic)
```

Table 3: Data summary

Name	dati_score_topic
Number of rows	447
Number of columns	14
Column type frequency:	
character	3
numeric	11
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Country	0	1	5	15	0	38	0
Region	0	1	4	35	0	447	0
Code	0	1	3	5	0	447	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Education	24	0.95	6.59	3.11	0.01	4.48	7.66	9.17	10.01
Jobs	15	0.97	6.47	2.59	0.01	4.78	7.17	8.37	10.01
Income	12	0.97	3.44	2.57	0.00	1.66	3.34	4.21	10.01
Safety	3	0.99	8.50	2.47	0.00	8.29	9.65	9.84	10.01
Health	0	1.00	5.84	2.76	0.01	3.79	6.36	8.17	10.00
Environment	0	1.00	6.58	2.60	0.01	5.17	7.11	8.46	10.00
Civic engagement	1	1.00	5.21	2.82	0.01	3.17	5.17	7.43	10.01
Accessiblity to services	0	1.00	6.09	2.28	0.01	4.76	6.50	7.94	10.00
Housing	3	0.99	4.50	2.95	0.01	1.74	4.30	6.80	10.01
Community	10	0.98	6.29	2.71	0.01	4.64	7.03	8.38	10.01
Life satisfaction	10	0.98	5.93	2.76	0.01	3.85	6.15	8.08	10.01

The means and medians for all variables are observed to be very similar. This suggests that the distributions are largely symmetrical, with the central tendency being well-represented by both measures.

The standard deviations across all variables are relatively low. This indicates a limited spread or dispersion of data points around their respective means, implying that observations within each variable are generally clustered closely together.

While missing values are present in the dataset, the completeness of observations is at least 90%. This level of data completeness is generally acceptable, though the impact of the missing data on specific analyses should be considered.

Specifically, two variables (Income and Safety) appear to exhibit asymmetry in their distributions. Further investigation, possibly through skewness statistics or visual aids like histograms and box plots, would be beneficial to quantify the direction and degree of this asymmetry

IMPUTAZIONE NA

```r
library(tidyverse)
```

```
## Warning: il pacchetto 'tidyverse' è stato creato con R versione 4.2.3

## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.2.3

## Warning: il pacchetto 'tibble' è stato creato con R versione 4.2.3

## Warning: il pacchetto 'tidyr' è stato creato con R versione 4.2.3

## Warning: il pacchetto 'readr' è stato creato con R versione 4.2.3

## Warning: il pacchetto 'purrr' è stato creato con R versione 4.2.3

## Warning: il pacchetto 'dplyr' è stato creato con R versione 4.2.3

## Warning: il pacchetto 'stringr' è stato creato con R versione 4.2.3

## Warning: il pacchetto 'forcats' è stato creato con R versione 4.2.3

## Warning: il pacchetto 'lubridate' è stato creato con R versione 4.2.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)

# 2. Impostazione per l'imputazione
# Identifica le colonne numeriche su cui applicare l'imputazione
# Escludi eventuali colonne ID o categoriche che non dovrebbero essere imputate
numeric_cols <- dati_score_topic %>%
  select(where(is.numeric)) %>%
  colnames()

# --- Passaggio 1: Imputazione della media per gruppo (Paese) ---
# Se un gruppo ha tutti NA per una variabile, la media del gruppo risulterà NA.
df_imputed_step1 <- dati_score_topic %>%
  group_by(Country) %>% # Raggruppa il dataframe per 'Country'
  mutate(across(all_of(numeric_cols), ~ {
    # Calcola la media del gruppo. Se tutti i valori nel gruppo sono NA, la media sarà NA.
    group_mean <- mean(., na.rm = TRUE)
    # Sostituisci NA con la media del gruppo. Se group_mean è NA, lascerà NA.
```

```
    replace_na(., group_mean)
  })) %>%
  ungroup() # Rimuovi il raggruppamento

print("\nDataset dopo l'imputazione con la media del Paese:")
```

```
## [1] "\nDataset dopo l'imputazione con la media del Paese:"
```

```
print(df_imputed_step1)
```

```
## # A tibble: 447 x 14
##    Country   Region       Code  Education  Jobs Income Safety Health Environment
##    <chr>     <chr>        <chr>     <dbl> <dbl>  <dbl>  <dbl>  <dbl>       <dbl>
##  1 Australia New South W~ AU1        8.51  8.23   5.68   9.81   9.47        7.56
##  2 Australia Victoria     AU2        8.51  8.42   4.60   9.76   9.95        8.31
##  3 Australia Queensland   AU3        8.19  8.35   5.02   9.78   9.21        9.30
##  4 Australia South Austr~ AU4        7.58  8.10   5.30   9.92   9.10        9.60
##  5 Australia Western Aus~ AU5        8.19  8.85   5.91   9.76   9.62        8.46
##  6 Australia Tasmania     AU6        7.23  8.07   4.94   9.86   8.26        9.00
##  7 Australia Northern Te~ AU7        8.08  8.49   7.19   9.70   6.27        8.81
##  8 Australia Australian ~ AU8        9.35  9.28  10.0    9.84   9.89        8.16
##  9 Austria   Burgenland   AT11       8.71  7.91   4.71   9.76   7.45        6.87
## 10 Austria   Lower Austr~ AT12       8.74  8.26   4.77   9.89   7.05        6.97
## # i 437 more rows
## # i 5 more variables: `Civic engagement` <dbl>,
## #   `Accessiblity to services` <dbl>, Housing <dbl>, Community <dbl>,
## #   `Life satisfaction` <dbl>
```

```
print("\nNA per colonna dopo Step 1 (media del Paese):")
```

```
## [1] "\nNA per colonna dopo Step 1 (media del Paese):"
```

```
print(colSums(is.na(df_imputed_step1)))
```

```
##                 Country                  Region                    Code
##                       0                       0                       0
##               Education                    Jobs                  Income
##                      12                       0                       0
##                  Safety                  Health             Environment
##                       0                       0                       0
##       Civic engagement Accessiblity to services                 Housing
##                       0                       0                       2
##               Community       Life satisfaction
##                       0                       0
```

```
# --- Passaggio 2: Imputazione della mediana globale per i rimanenti NA ---
# Questi NA sono quelli che non sono stati imputati al Passaggio 1 perché interi gruppi erano NA.
df_final_imputed <- df_imputed_step1 %>%
  mutate(across(all_of(numeric_cols), ~ {
    # Calcola la mediana globale per la colonna.
```

```
    # Questo è necessario perché 'group_mean' potrebbe essere ancora NA
    # se l'intero gruppo aveva solo NA per quella variabile.
    global_median <- median(df_imputed_step1[[cur_column()]], na.rm = TRUE)
    # Sostituisci i rimanenti NA (quelli dai gruppi con tutti NA) con la mediana globale
    replace_na(., global_median)
  }))

print("\nDataset Finale dopo Imputazione (Media del Paese, poi Mediana Globale):")
```

```
## [1] "\nDataset Finale dopo Imputazione (Media del Paese, poi Mediana Globale):"
```

```
print(df_final_imputed)
```

```
## # A tibble: 447 x 14
##    Country   Region       Code  Education  Jobs Income Safety Health Environment
##    <chr>     <chr>        <chr>     <dbl> <dbl>  <dbl>  <dbl>  <dbl>       <dbl>
##  1 Australia New South W~ AU1        8.51  8.23   5.68   9.81   9.47        7.56
##  2 Australia Victoria     AU2        8.51  8.42   4.60   9.76   9.95        8.31
##  3 Australia Queensland   AU3        8.19  8.35   5.02   9.78   9.21        9.30
##  4 Australia South Austr~ AU4        7.58  8.10   5.30   9.92   9.10        9.60
##  5 Australia Western Aus~ AU5        8.19  8.85   5.91   9.76   9.62        8.46
##  6 Australia Tasmania     AU6        7.23  8.07   4.94   9.86   8.26        9.00
##  7 Australia Northern Te~ AU7        8.08  8.49   7.19   9.70   6.27        8.81
##  8 Australia Australian ~ AU8        9.35  9.28  10.0    9.84   9.89        8.16
##  9 Austria   Burgenland   AT11       8.71  7.91   4.71   9.76   7.45        6.87
## 10 Austria   Lower Austr~ AT12       8.74  8.26   4.77   9.89   7.05        6.97
## # i 437 more rows
## # i 5 more variables: 'Civic engagement' <dbl>,
## #   'Accessiblity to services' <dbl>, Housing <dbl>, Community <dbl>,
## #   'Life satisfaction' <dbl>
```

```
print("\nNA per colonna dopo l'imputazione finale:")
```

```
## [1] "\nNA per colonna dopo l'imputazione finale:"
```

```
print(colSums(is.na(df_final_imputed)))
```

```
##                   Country                    Region                      Code
##                         0                         0                         0
##                 Education                      Jobs                    Income
##                         0                         0                         0
##                    Safety                    Health               Environment
##                         0                         0                         0
##         Civic engagement Accessiblity to services                   Housing
##                         0                         0                         0
##                 Community        Life satisfaction
##                         0                         0
```

Check if there are missing values:

```
last_score_region <- df_final_imputed
skim_without_charts(last_score_region)
```

Table 6: Data summary

Name	last_score_region
Number of rows	447
Number of columns	14
Column type frequency:	
character	3
numeric	11
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Country	0	1	5	15	0	38	0
Region	0	1	4	35	0	447	0
Code	0	1	3	5	0	447	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Education	0	1	6.56	3.07	0.01	4.36	7.61	9.13	10.01
Jobs	0	1	6.40	2.61	0.01	4.66	7.10	8.36	10.01
Income	0	1	3.38	2.57	0.00	1.28	3.25	4.15	10.01
Safety	0	1	8.48	2.48	0.00	8.27	9.65	9.84	10.01
Health	0	1	5.84	2.76	0.01	3.79	6.36	8.17	10.00
Environment	0	1	6.58	2.60	0.01	5.17	7.11	8.46	10.00
Civic engagement	0	1	5.20	2.83	0.01	3.17	5.16	7.41	10.01
Accessiblity to services	0	1	6.09	2.28	0.01	4.76	6.50	7.94	10.00
Housing	0	1	4.50	2.94	0.01	1.74	4.33	6.80	10.01
Community	0	1	6.32	2.70	0.01	4.65	7.03	8.38	10.01
Life satisfaction	0	1	5.94	2.74	0.01	3.85	6.15	8.08	10.01

Now we observe the variables distributions:

```
library(ggplot2)
library(dplyr)
library(purrr)
library(patchwork)
```

```
## Warning: il pacchetto 'patchwork' è stato creato con R versione 4.2.3
```

```r
df <- last_score_region

# --- 2. Selects only numeric variables ---

numeric_vars <- df %>%
  select(where(is.numeric)) %>%
  colnames()

print(paste("\nNumerical variables identified for histograms:", length(numeric_vars)))
```

```
## [1] "\nNumerical variables identified for histograms: 11"
```

```r
print(numeric_vars)
```

```
##  [1] "Education"                "Jobs"
##  [3] "Income"                   "Safety"
##  [5] "Health"                   "Environment"
##  [7] "Civic engagement"         "Accessiblity to services"
##  [9] "Housing"                  "Community"
## [11] "Life satisfaction"
```

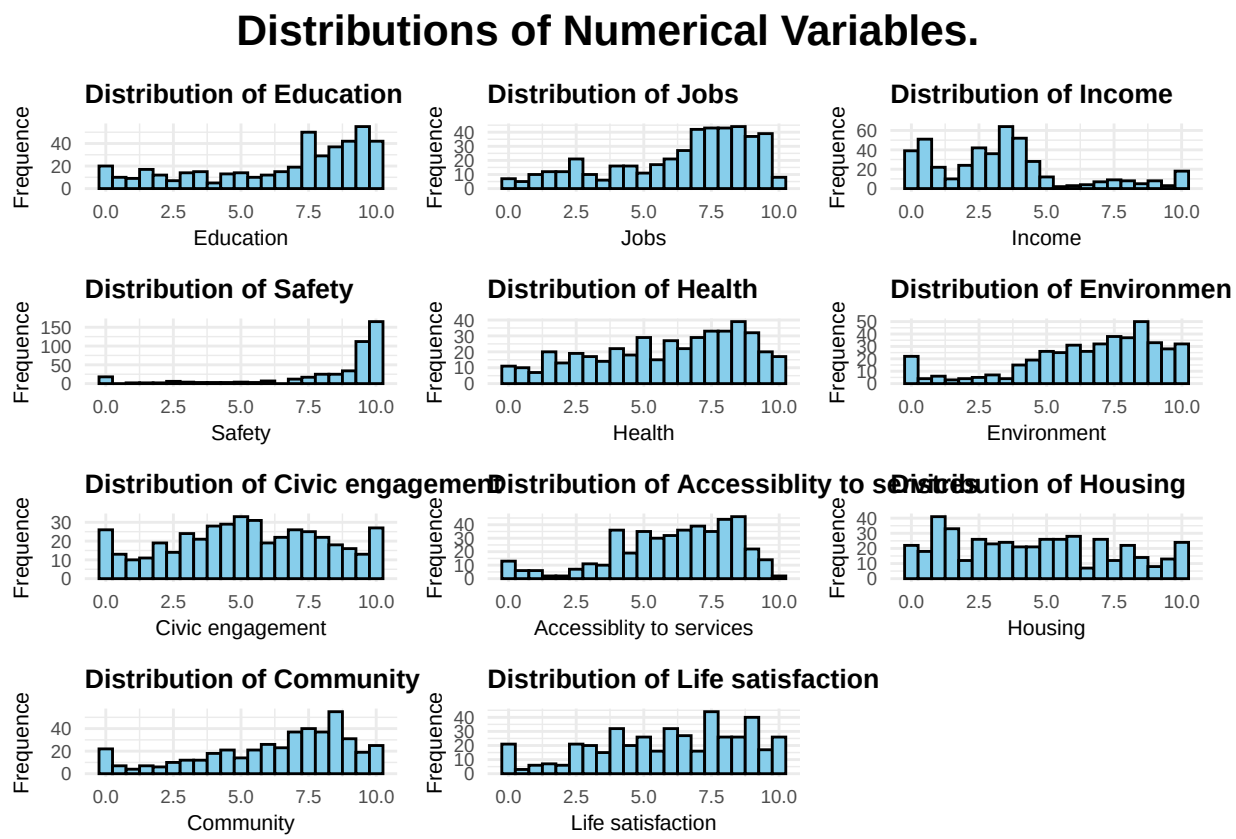
```r
histogram_list <- map(numeric_vars, ~ {
  ggplot(df, aes(x = .data[[.x]])) +
    geom_histogram(binwidth = (max(df[[.x]], na.rm = TRUE) - min(df[[.x]], na.rm = TRUE)) / 20,
                                    fill = "skyblue", color = "black") +
    labs(title = paste("Distribution of", .x), x = .x, y = "Frequence") +
    theme_minimal() + # Tema pulito
    theme(plot.title = element_text(size = 10, face = "bold"),
          axis.title = element_text(size = 8),
          axis.text = element_text(size = 7))
})

# --- 4. Histograms ---

num_plots <- length(histogram_list)
cols_per_row <- 3
rows_needed <- ceiling(num_plots / cols_per_row)


combined_plot <- wrap_plots(histogram_list, ncol = cols_per_row, nrow = rows_needed) +
  plot_annotation(
    title = "Distributions of Numerical Variables.",
    theme = theme(plot.title = element_text(size = 16, face = "bold", hjust = 0.5))
  )

# Show the combined plot
print(combined_plot)
```



In general, most of the distributions of the numerical indicators show a tendency to be left-skewed. This means that the majority of observations are concentrated towards the higher values of the range, with a longer "tail" towards the lower values. This is a positive sign for well-being indicators, as it suggests that most regions are performing relatively well in these areas. Many of the distributions show a shape resembling a unimodal distribution, with a peak (mode) towards the upper end of the scale. SAFETY: This is the most distinctive distribution. It is strongly left-skewed (negatively skewed) with a very pronounced and dominant peak in the last one or two bins (towards the value 10). This indicates that the vast majority of regions have extremely high safety scores, close to the maximum. INCOME: Although it maintains some left-skewness, the income distribution shows a less sharp peak compared to Safety and a wider spread across the entire range. There are several significant bars across the entire scale, indicating greater variability in income scores among regions compared to other indicators like Safety. CIVIC ENGAGEMENT: Unlike most others, this distribution appears to be more bimodal or have a more pronounced central plateau. There are two noticeable peaks, one towards the lower values (between 2 and 3) and another towards the middle-to-high end of the scale (between 6 and 8). This could suggest the existence of two distinct groups of regions: those with low civic engagement and those with moderate-to-high engagement, with fewer regions in between.

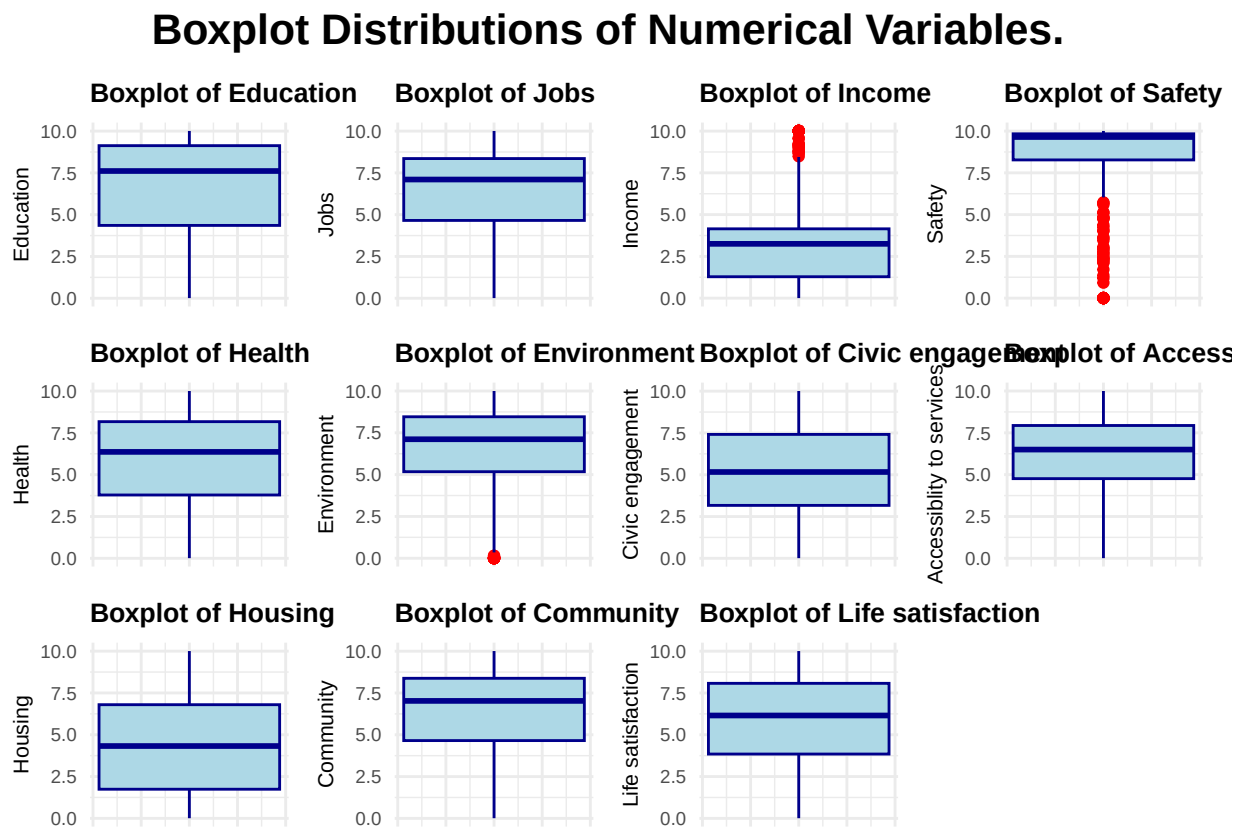
```
boxplot_list <- map(numeric_vars, ~ {
  ggplot(df, aes(y = .data[[.x]])) +
    geom_boxplot(fill = "lightblue", color = "darkblue", outlier.colour = "red") +
    labs(title = paste("Boxplot of", .x), y = .x) +
    theme_minimal() +
    theme(plot.title = element_text(size = 10, face = "bold"),
          axis.title.y = element_text(size = 8),
          axis.text.y = element_text(size = 7),
          axis.text.x = element_blank(),
          axis.ticks.x = element_blank())
```

```
})

# --- 4. boxplot  ---
num_plots <- length(boxplot_list)
cols_per_row <- 4
rows_needed <- ceiling(num_plots / cols_per_row)

combined_boxplot_plot <- wrap_plots(boxplot_list, ncol = cols_per_row, nrow = rows_needed) +
  plot_annotation(
    title = "Boxplot Distributions of Numerical Variables.",
    theme = theme(plot.title = element_text(size = 16, face = "bold", hjust = 0.5))
  )

# Show the combined plot
print(combined_boxplot_plot)
```



From a general perspective, most variables (e.g., Education, Jobs, Health, Environment, Accessibility to services, Housing, Community, Life satisfaction) show medians clustered in the upper half of the 0-10 scale, typically between 5.0 and 8.0, with relatively few prominent outliers. This suggests that for a majority of these indicators, the dataset's observations generally lean towards higher scores. However, a closer look at "Income" and "Safety" reveals distinct distributional characteristics.

SAFETY: This boxplot strongly confirms the severe left-skewness observed in the histogram for Safety. The vast majority of regions have very high safety scores, with little variability among them. The outliers indicate that there are a number of regions with unusually low safety scores, which deviate significantly from the norm.

INCOME: This indicates that while the majority of regions (middle 50%) have relatively low-to-moderate income scores, there are a substantial number of regions that are performing exceptionally well in terms of income, acting as outliers. This distribution shows a strong positive skewness (skewed to the right) or a very long tail of high-income regions, which aligns with the "Income" histogram having a wider spread but with values clustered on the lower side for the bulk of the data, and some high values on the far right.

The others are normal.

```
cov_var <- round(cov(last_score_region[,4:14], use = "complete.obs"), 3)
cor_matrix <- round(cor(last_score_region[, 4:14], use = "complete.obs"), 3)
cor_matrix
```

```
##                           Education  Jobs Income Safety Health Environment
## Education                     1.000 0.651  0.682  0.563  0.046       0.448
## Jobs                          0.651 1.000  0.533  0.381  0.026       0.394
## Income                        0.682 0.533  1.000  0.378  0.156       0.556
## Safety                        0.563 0.381  0.378  1.000  0.352       0.186
## Health                        0.046 0.026  0.156  0.352  1.000       0.078
## Environment                   0.448 0.394  0.556  0.186  0.078       1.000
## Civic engagement              0.067 0.144  0.183  0.262  0.298       0.095
## Accessiblity to services      0.616 0.521  0.602  0.614  0.290       0.275
## Housing                       0.604 0.541  0.771  0.425  0.209       0.653
## Community                     0.501 0.407  0.456  0.343  0.200       0.505
## Life satisfaction             0.538 0.564  0.554  0.269  0.245       0.618
##                           Civic engagement Accessiblity to services Housing
## Education                            0.067                     0.616   0.604
## Jobs                                 0.144                     0.521   0.541
## Income                               0.183                     0.602   0.771
## Safety                               0.262                     0.614   0.425
## Health                               0.298                     0.290   0.209
## Environment                          0.095                     0.275   0.653
## Civic engagement                     1.000                     0.155   0.371
## Accessiblity to services             0.155                     1.000   0.564
## Housing                              0.371                     0.564   1.000
## Community                            0.203                     0.414   0.526
## Life satisfaction                    0.163                     0.467   0.585
##                           Community Life satisfaction
## Education                     0.501             0.538
## Jobs                          0.407             0.564
## Income                        0.456             0.554
## Safety                        0.343             0.269
## Health                        0.200             0.245
## Environment                   0.505             0.618
## Civic engagement              0.203             0.163
## Accessiblity to services      0.414             0.467
## Housing                       0.526             0.585
## Community                     1.000             0.634
## Life satisfaction             0.634             1.000
```

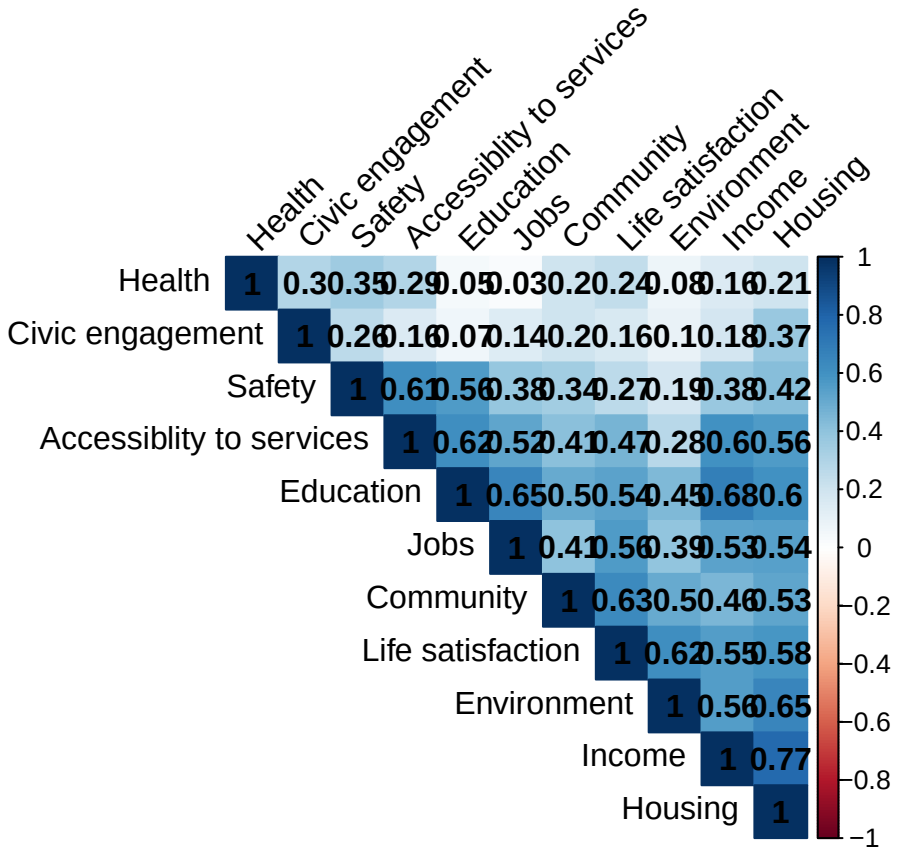
All the correlation are positive and most of them are positively moderate correlated. The hightest correlation is 0.765 between Housing and Income. The 5 variables that are most correlated with our target variable (Life satisfaction) are: Community, Environment, Housing, Jobs and Income.

```
library(corrplot)
```

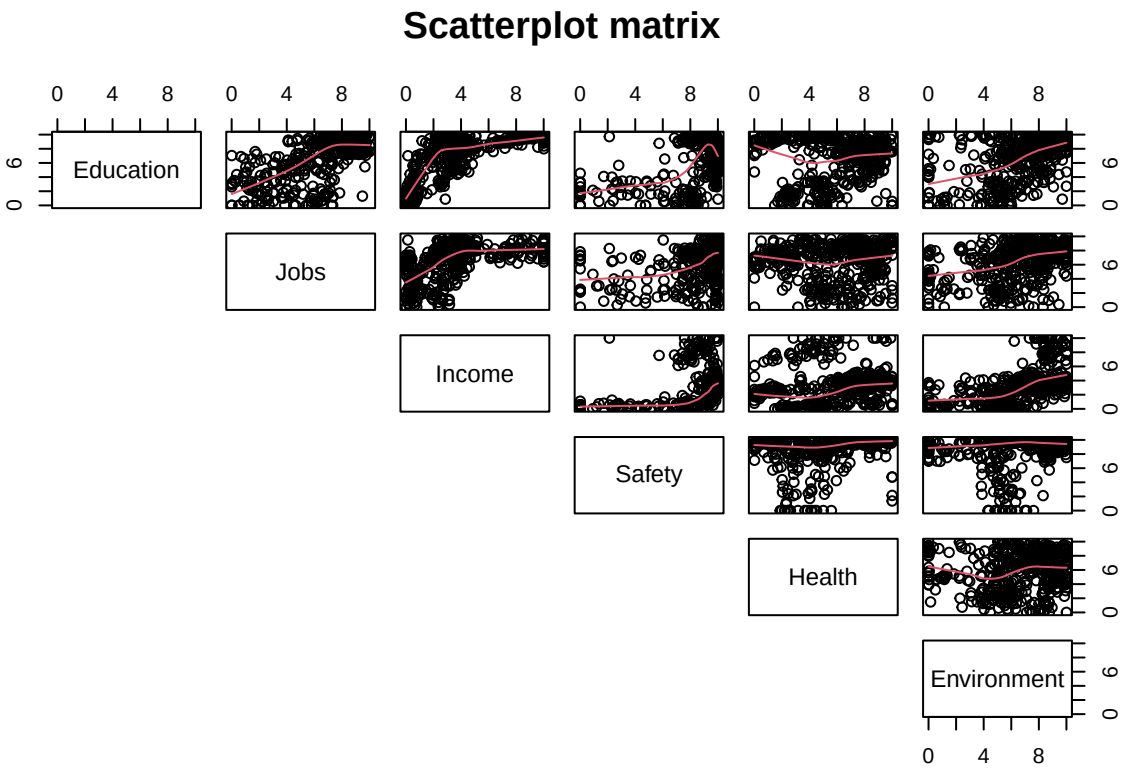
```
## Warning: il pacchetto 'corrplot' è stato creato con R versione 4.2.3
```

```
## corrplot 0.92 loaded
```

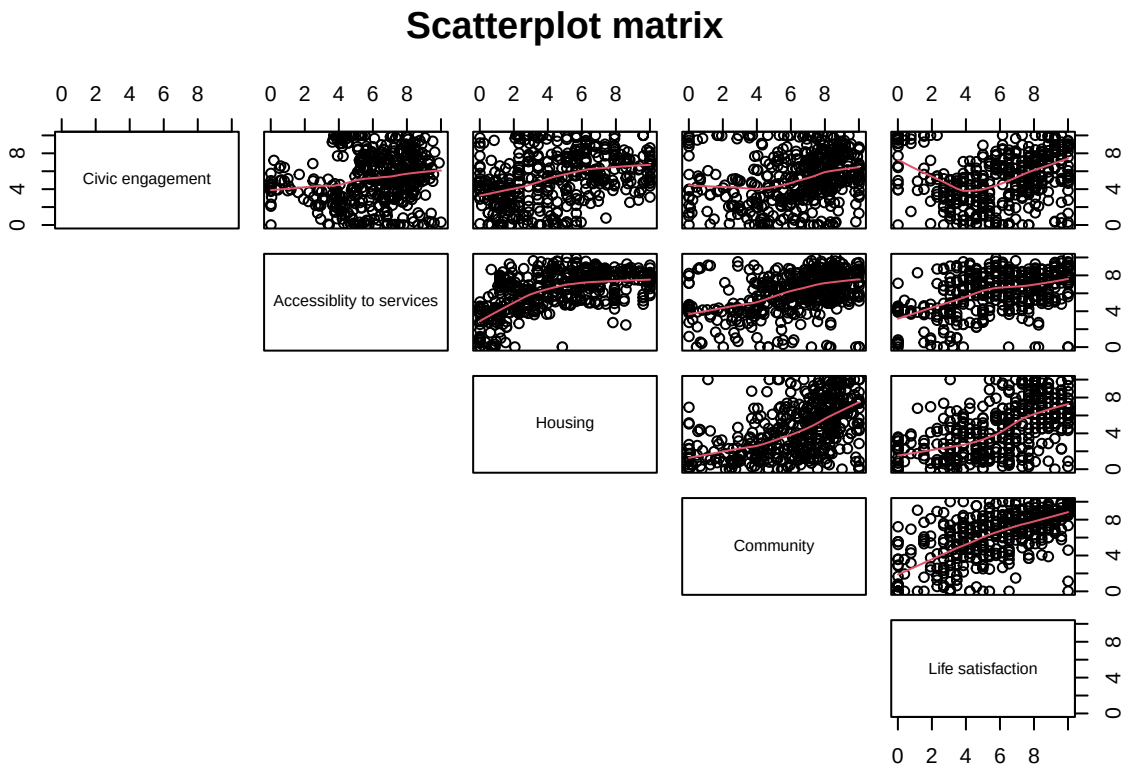
```
corrplot(cor_matrix,
         method = "color",
         type = "upper",
         order = "hclust",
         tl.col = "black",
         tl.srt = 45,
         addCoef.col = "black",
         col = COL2("RdBu", 200) )
```




```
pairs(last_score_region[,4:9],
 panel = panel.smooth,
lower.panel = NULL,
main = "Scatterplot matrix")
```




```
pairs(last_score_region[,10:14],
 panel = panel.smooth,
lower.panel = NULL,
main = "Scatterplot matrix")
```



There is a consistent positive correlation among all variables.

DATASET LAST SCORE COUNTRY

```
numeric_cols_for_aggregation <- last_score_region %>%
  select(where(is.numeric)) %>% # Seleziona tutte le colonne numeriche
  colnames()

last_score_country <- last_score_region %>%
  group_by(Country) %>% # Raggruppa il DataFrame per la colonna 'Country'
  summarise(
    # Calcola la media per ogni colonna numerica selezionata
    # .across() è utile per applicare la stessa funzione a più colonne
    # .names = "{.col}_mean" è un'opzione per rinominare le nuove colonne (es. Education_mean)
    # na.rm = TRUE è cruciale per ignorare i valori NA nel calcolo della media
    across(all_of(numeric_cols_for_aggregation), mean, na.rm = TRUE, .names = "{.col}_mean")
  ) %>%
  ungroup() # Rimuovi il raggruppamento per evitare problemi in operazioni future
```

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(...)`.
## i In group 1: `Country = "Australia"`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
```

```
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

```
print("\nDataset Raggruppato per Paese (Medie):")
```

```
## [1] "\nDataset Raggruppato per Paese (Medie):"
```

```
head(last_score_country)
```

```
## # A tibble: 6 x 12
##   Country   Education_mean Jobs_mean Income_mean Safety_mean Health_mean
##   <chr>              <dbl>     <dbl>       <dbl>       <dbl>       <dbl>
## 1 Australia           8.21      8.47       6.08         9.80        8.97
## 2 Austria             8.58      7.99       4.55         9.89        7.64
## 3 Belgium             7.48      5.44       3.93         9.61        6.58
## 4 Canada              9.00      7.10       4.64         9.19        6.18
## 5 Chile               5.65      4.94       0.803        9.13        8.31
## 6 Colombia            3.24      2.41       0.564        3.22        4.73
## # i 6 more variables: Environment_mean <dbl>, `Civic engagement_mean` <dbl>,
## #   `Accessiblity to services_mean` <dbl>, Housing_mean <dbl>,
## #   Community_mean <dbl>, `Life satisfaction_mean` <dbl>
```

```
library(ggplot2)
library(dplyr)
#install.packages("sf") # Per dati spaziali (geometrie)
#install.packages("rnaturalearth") # Per ottenere dati geografici del mondo
#install.packages("rnaturalearthdata") # Dati geografici più dettagliati per rnaturalearth
```

```
library(ggplot2)
library(dplyr)
library(sf) # Simple Features for R
```

```
## Warning: il pacchetto 'sf' è stato creato con R versione 4.2.3
```

```
## Linking to GEOS 3.9.3, GDAL 3.5.2, PROJ 8.2.1; sf_use_s2() is TRUE
```

```
library(rnaturalearth) # Per ottenere i dati geografici dei paesi
```

```
## Warning: il pacchetto 'rnaturalearth' è stato creato con R versione 4.2.3
```

```
library(rnaturalearthdata) # Dati aggiuntivi per rnaturalearth
```

```
## Warning: il pacchetto 'rnaturalearthdata' è stato creato con R versione 4.2.3
```

```
##
## Caricamento pacchetto: 'rnaturalearthdata'

## Il seguente oggetto è mascherato da 'package:rnaturalearth':
##
##     countries110
```

```
# --- 2. Ottieni i dati geografici del mondo ---
# Usiamo ne_countries per ottenere le geometrie dei paesi
world <- ne_countries(scale = "medium", returnclass = "sf")

# Visualizza alcune righe dei dati geografici per capire le colonne disponibili
# In particolare, cerca la colonna che contiene i nomi dei paesi (spesso 'name' o 'sovereignt')
print("\nPrime righe del dataset geografico 'world':")
```

```
## [1] "\nPrime righe del dataset geografico 'world':"
```

```
print(head(world %>% select(name, sovereignt, geometry)))
```

```
## Simple feature collection with 6 features and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -73.36621 ymin: -22.40205 xmax: 109.4449 ymax: 41.9062
## Geodetic CRS:  WGS 84
##        name sovereignt                       geometry
## 1  Zimbabwe   Zimbabwe MULTIPOLYGON (((31.28789 -2...
## 2    Zambia     Zambia MULTIPOLYGON (((30.39609 -1...
## 3     Yemen      Yemen MULTIPOLYGON (((53.08564 16...
## 4   Vietnam    Vietnam MULTIPOLYGON (((104.064 10....
## 5 Venezuela  Venezuela MULTIPOLYGON (((-60.82119 9...
## 6   Vatican    Vatican MULTIPOLYGON (((12.43916 41...
```

```
# --- 3. Unisci il tuo dataset con i dati geografici ---
# L'unione (join) è la parte più critica. Devi assicurarti che la colonna 'Country'
# nel tuo dataset corrisponda a una colonna di nomi di paesi nel dataset 'world'.
# Spesso 'name' o 'sovereignt' in 'world' funziona bene.

# Effettua un left_join. Manterremo tutte le geometrie dei paesi,
# e aggiungeremo i tuoi dati di Life_satisfaction dove i paesi corrispondono.
# I paesi nel dataset geografico che non sono nel tuo dataset avranno NA per Life_satisfaction.
world_data <- left_join(world, last_score_country, by = c("name" = "Country"))

# Puoi verificare quanti paesi hanno trovato corrispondenza:
print(paste("\nNumero di paesi nel dataset geografico dopo l'unione:", nrow(world_data)))
```

```
## [1] "\nNumero di paesi nel dataset geografico dopo l'unione: 242"
```

```
print(paste("Numero di paesi con dati di 'Life_satisfaction' mappati:", sum(!is.na(world_data$`Life sati
```

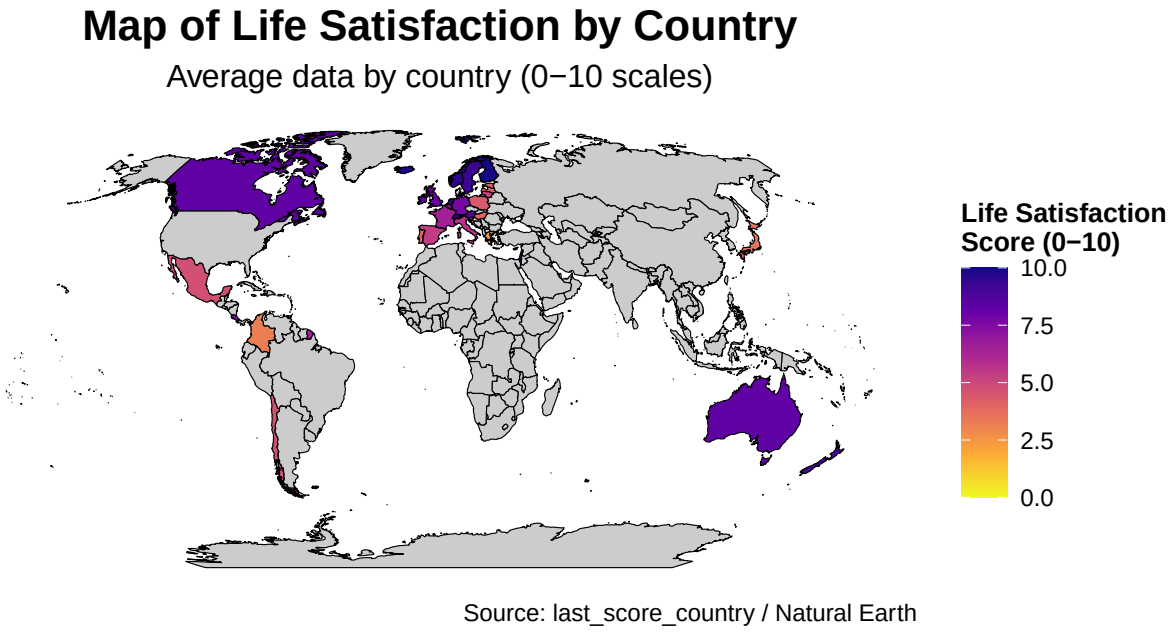
```
## [1] "Numero di paesi con dati di 'Life_satisfaction' mappati: 33"
```

```r
print(paste("Paesi nel tuo dataset non trovati nel GeoJSON (potrebbero essere presenti ma con nomi dive
            paste(last_score_country$Country[!last_score_country$Country %in% world_data$name], collaps
```

```
## [1] "Paesi nel tuo dataset non trovati nel GeoJSON (potrebbero essere presenti ma con nomi diversi):
```

```r
# --- 4. Crea la mappa coropletica con ggplot2 ---

ggplot(data = world_data) +
  geom_sf(aes(fill = `Life satisfaction_mean`), color = "black", size = 0.1) + # color/size per i bordi
  scale_fill_viridis_c(option = "plasma", direction = -1, # Scala di colori continua (viridis, plasma,
                       na.value = "grey80", # Colore per i paesi con dati mancanti
                       limits = c(0, 10), # Imposta il range dei colori da 0 a 10
                       name = "Life Satisfaction\nScore (0-10)") + # Etichetta della legenda
  coord_sf(crs = "+proj=robin") + # Proiezione della mappa (Robinson è comune per mappe mondiali)
  labs(
    title = "Map of Life Satisfaction by Country",
    subtitle = "Average data by country (0-10 scales)",
    caption = "Source: last_score_country / Natural Earth"
  ) +
  theme_minimal() + # Tema pulito
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    plot.subtitle = element_text(hjust = 0.5, size = 12),
    legend.position = "right", # Posizione della legenda
    legend.title = element_text(size = 10, face = "bold"),
    legend.text = element_text(size = 9),
    panel.grid.major = element_line(color = "transparent"), # Rimuove le griglie
    panel.grid.minor = element_line(color = "transparent"),
    axis.text = element_blank(), # Rimuove le etichette degli assi (lat/long)
    axis.ticks = element_blank() # Rimuove i tick degli assi
  )
```



The choropleth map visually represents "Life Satisfaction Scores" across various countries, using a color gradient from yellow (low satisfaction) to purple (high satisfaction). Countries for which data is not available appear in grey.

Several European countries, particularly in Scandinavia (e.g., Denmark, Finland, Norway), show high levels of life satisfaction, indicated by darker purple shades. Australia and New Zeland also appear with high scores.

Countries in regions like South America (e.g., Chile, parts of South America where data is present) and parts of Eastern Europe tend to display lower life satisfaction, represented by lighter orange/pink colors.

The graph below illustrates the observations previously made:

```
# (Necessita dei pacchetti ggplot2, dplyr, sf, rnaturalearth, rnaturalearthdata come prima)

# 1. Ottieni i dati geografici e calcola i centroidi per i punti delle bolle
world_centroids <- ne_countries(scale = "medium", returnclass = "sf") %>%
  st_centroid() # Calcola il centroide di ogni geometria del paese
```

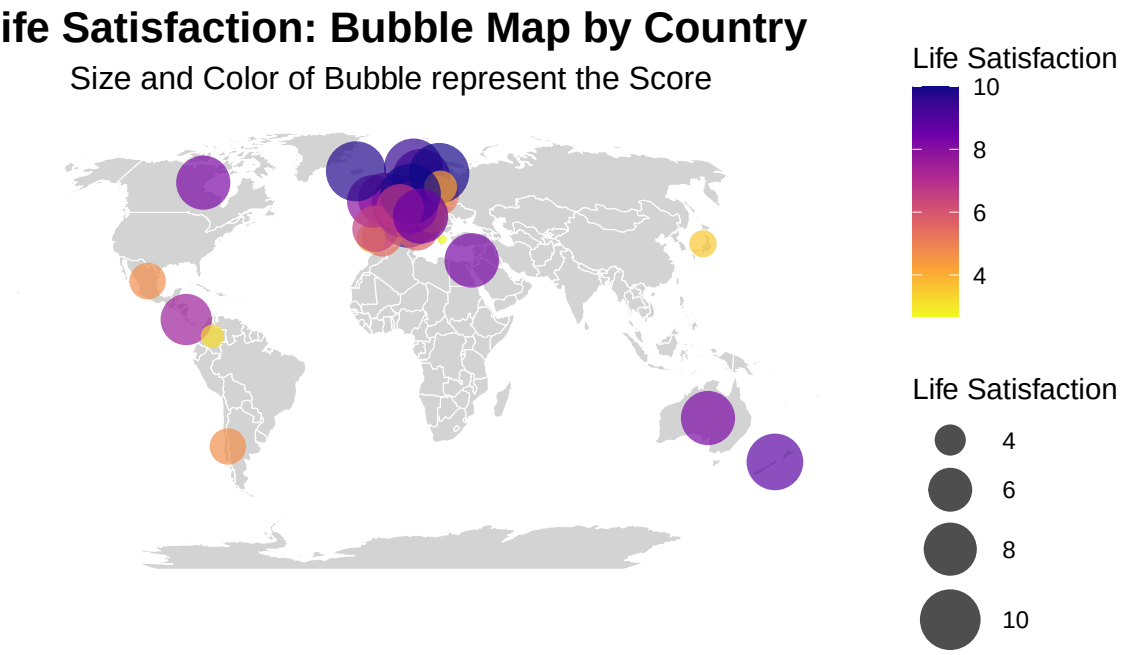
```
## Warning: st_centroid assumes attributes are constant over geometries
```

```
# 2. Unisci il tuo dataset con i centroidi
world_data_centroids <- left_join(world_centroids, last_score_country, by = c("name" = "Country"))

# 3. Ottieni le geometrie dei paesi (per lo sfondo della mappa)
world_map_background <- ne_countries(scale = "medium", returnclass = "sf")
```

```r
# 4. Crea la mappa a bolle
ggplot() +
  geom_sf(data = world_map_background, fill = "lightgrey", color = "white", size = 0.1) + # Mappa di ba
  geom_sf(data = world_data_centroids, aes(size = `Life satisfaction_mean`, color = `Life satisfaction_m
  scale_size_continuous(range = c(1, 10), name = "Life Satisfaction") + # Range di dimensione delle bol
  scale_color_viridis_c(option = "plasma", direction = -1, name = "Life Satisfaction") + # Colore delle
  coord_sf(crs = "+proj=robin") +
  labs(title = "Life Satisfaction: Bubble Map by Country",
       subtitle = "Size and Color of Bubble represent the Score") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    plot.subtitle = element_text(hjust = 0.5, size = 12),
    legend.position = "right",
    axis.text = element_blank(),
    axis.ticks = element_blank(),
    panel.grid.major = element_line(color = "transparent"),
    panel.grid.minor = element_line(color = "transparent")
  )
```

```
## Warning: Removed 209 rows containing missing values (`geom_sf()`)
```




```r
library(dplyr)

df_ordered_by_lifesat <- last_score_country %>%
```

```
  arrange(`Life satisfaction_mean`)

df_ordered_by_lifesat[1:3,]
```

```
## # A tibble: 3 x 12
##   Country  Education_mean Jobs_mean Income_mean Safety_mean Health_mean
##   <chr>             <dbl>     <dbl>       <dbl>       <dbl>       <dbl>
## 1 Türkiye           0.816      2.74       0.513        8.01        5.13
## 2 Greece            6.64       2.08       2.31         9.86        7.26
## 3 Colombia          3.24       2.41       0.564        3.22        4.73
## # i 6 more variables: Environment_mean <dbl>, `Civic engagement_mean` <dbl>,
## #   `Accessiblity to services_mean` <dbl>, Housing_mean <dbl>,
## #   Community_mean <dbl>, `Life satisfaction_mean` <dbl>
```

```
df_ordered_by_lifesat[36:38,]
```

```
## # A tibble: 3 x 12
##   Country Education_mean Jobs_mean Income_mean Safety_mean Health_mean
##   <chr>            <dbl>     <dbl>       <dbl>       <dbl>       <dbl>
## 1 Iceland           7.61      8.84        3.46        9.83        8.32
## 2 Finland           8.92      7.74        3.75        9.72        8.18
## 3 Denmark           7.71      8.45        3.43        9.79        7.34
## # i 6 more variables: Environment_mean <dbl>, `Civic engagement_mean` <dbl>,
## #   `Accessiblity to services_mean` <dbl>, Housing_mean <dbl>,
## #   Community_mean <dbl>, `Life satisfaction_mean` <dbl>
```

An analysis of life satisfaction reveals that the three lowest-ranking countries are Turkey, Greece, and Colombia. Consistent with our expectations, the variables most highly correlated with life satisfaction (especially income) exhibit notably low values for these nations. In stark contrast, Iceland, Finland, and Denmark emerge as the top three countries in terms of life satisfaction. For these high-ranking countries, the values of the most correlated variables are very high. These findings suggest a potential link to differing political wellness across these countries.