Regional Well Being

Akash Mittal, Maged Saeed Abdo Mostafa Kharshom, Precious Prince, Franco Reinaldo Bonifacini

2024-12-18

Abstract

This project aims to **estimate the effect of various factors such as health, education, and income among others, on the life satisfaction of people in different regions from different countries**. For this, we carried out a robust approach with MEMs & Robust Estimators. Also, we performed a robust clustering of various factors into different clusters of life satisfacton.

Introduction

For this project, we utilized the OECD database focusing on **indicators and life-satisfaction scores across various regions within different countries**. This comprehensive dataset captures diverse dimensions of well-being, such as education, employment, health, environment, and social support, which together contribute to a region's overall quality of life. The indicators allow for meaningful comparisons between regions, highlighting disparities and trends. Although the data reflects different years depending on the country, for consistency we adopted the **latest available data**.

It is essential to acknowledge that while some variability exists in the timing of the data, we assumed that any changes within a year or two would be minimal and unlikely to significantly alter policies or life-satisfaction scores. This approach enables us to draw relevant conclusions about the factors influencing regional well-being. By examining key metrics such as income, safety, health, and life satisfaction, this report aims to present a clear picture of well-being across OECD regions and highlight patterns that can inform future policies and initiatives.

Original Dataset

To execute this project, a dataset containing the different metrics regarding well-being was downloaded and linked to a variable named **df_wb**.

The variable **df_wb** contains 447 tuples and 28 columns (25 are the attributes to analyze).

```
## [1] 447  28
```

The variables included are the following:

1. **Country:** Includes the name of all the countries included in the results.
2. **Region:** Includes the name of all the cities from the countries included in the results.
3. **Code:** Code associated to each pair country-city.
4. **Population.with.at.least.secondary.education.(%):** Percentage of the population completing secondary education.
5. **Employment rate (%):** Percentage of the working-age population employed.
6. **Unemploy-ment rate (%):** Percentage of people without jobs actively seeking work.
7. **Household disposable income per capita (USD PPP):** Average income available per person, in USD.
8. **Homicide rate (per 100k):** Number of homicides per 100.000 people.
9. **Mortality rate (per 1k):** Deaths per 1,000 people annually.
10. **Life expectancy:** Average expected lifespan (years).
11. **Air pollution (level of PM2.5, µg/m^3):** Fine particulate air pollution levels.
12. **Voter turnout (%):** Share of voters participating in elections.
13. **Broadband access (% of household):** Percentage of households with internet access.
14. **Internet download speed 2021-Q4 (%):** Internet speed growth/decline in 2021-Q4.
15. **Number of rooms per person:** Average living space per person.
16. **Perceived social network support (%):** Percentage of people with available social support.
17. **Self assessment of life satisfaction (0-10):** Subjective rating of overall happiness.
18. **Education (0-10):** Regional score for education.
19. **Jobs (0-10):** Score based on employment indicators.
20. **Income (0-10):** Score for household income.
21. **Safety (0-10):** Regional score for personal safety.
22. **Health (0-10):** Score for health indicators.
23. **Environment (0-10):** Score for environmental quality.
24. **Civic engagement (0-10):** Score for public participation.
25. **Accessibility to services (0-10):** Availability of public services.
26. **Housing (0-10):** Score for housing quality and affordability.
27. **Community (0-10):** Score for social cohesion.
28. **Life satisfaction (0-10):** Overall happiness score.

The value type of each attribute are the following:

```
##                                          Country
##                                      "character"
##                                           Region
##                                      "character"
##                                             Code
##                                      "character"
## Population.with.at.least.secondary.education.(%)
##                                      "character"
##                              Employment.rate.(%)
##                                      "character"
##                           Unemploy-ment.rate.(%)
##                                      "character"
## Household.disposable.income.per.capita.(USD.PPP)
##                                      "character"
##                         Homicide.rate.(per.100k)
##                                      "character"
##                          Mortality.rate.(per.1k)
##                                        "numeric"
##                                  Life.expectancy
##                                      "character"
##            Air.pollution.(level.of.PM2.5,.µg/m³)
##                                        "numeric"
##                                Voter.turnout.(%)
##                                      "character"
##                Broadband.access.(%.of.household)
##                                      "character"
##              Internet.download.speed.2021-Q4.(%)
##                                      "character"
##                       Number.of.rooms.per.person
##                                      "character"
##             Perceived.social.network.support.(%)
##                                      "character"
##      Self.assessment.of.life.satisfaction.(0-10)
##                                      "character"
##                                 Education.(0-10)
##                                      "character"
##                                      Jobs.(0-10)
##                                      "character"
##                                    Income.(0-10)
##                                      "character"
##                                    Safety.(0-10)
##                                      "character"
##                                    Health.(0-10)
##                                        "numeric"
##                               Environment.(0-10)
##                                        "numeric"
##                          Civic.engagement.(0-10)
##                                      "character"
##                  Accessiblity.to.services.(0-10)
##                                        "numeric"
##                                   Housing.(0-10)
##                                      "character"
##                                 Community.(0-10)
##                                      "character"
##                         Life.satisfaction.(0-10)
##                                      "character"
```

Data Manipulation and EDA

Null values

Before analyzing the hypothesis and attributes, a check on the data structure was conducted to prevent potential errors in the future. This involved examining both data types and null values.

First we checked the **null values** to determine their significance and understand which is the best action to take regaridng this matter. After checking that there were no null values, but yet we couldn't perform some calculations on the attributes, we did a more detailed analysis to realize that there were null values which were replaced by the **character ".."**.

```
##                                                   wb_missing_values
## Country                                                   0.0000000
## Region                                                   0.0000000
## Code                                                     0.0000000
## Population.with.at.least.secondary.education.(%)         5.3691275
## Employment.rate.(%)                                      3.3557047
## Unemploy-ment.rate.(%)                                   3.5794183
## Household.disposable.income.per.capita.(USD.PPP)         2.6845638
## Homicide.rate.(per.100k)                                 0.6711409
## Mortality.rate.(per.1k)                                  0.0000000
## Life.expectancy                                          1.7897092
## Air.pollution.(level.of.PM2.5,.µg/m³)                    0.0000000
## Voter.turnout.(%)                                        0.2237136
## Broadband.access.(%.of.household)                        1.3422819
## Internet.download.speed.2021-Q4.(%)                      0.6711409
## Number.of.rooms.per.person                               0.6711409
## Perceived.social.network.support.(%)                     2.2371365
## Self.assessment.of.life.satisfaction.(0-10)              2.2371365
## Education.(0-10)                                         5.3691275
## Jobs.(0-10)                                              3.3557047
## Income.(0-10)                                            2.6845638
## Safety.(0-10)                                            0.6711409
## Health.(0-10)                                            0.0000000
## Environment.(0-10)                                       0.0000000
## Civic.engagement.(0-10)                                  0.2237136
## Accessiblity.to.services.(0-10)                          0.0000000
## Housing.(0-10)                                           0.6711409
## Community.(0-10)                                         2.2371365
## Life.satisfaction.(0-10)                                 2.2371365
```

Taking this into account, we concluded that there were no significance level of null values (in the form of "..."), so for the moment we decided to keep all the information and look for further actions regarding null values.

```
##        Country n_cities n_missing      p_na
## 1       Canada       13         3  23.07692
## 2        Chile       16         1   6.25000
## 3     Colombia       33         9  27.27273
## 4   Costa Rica        6         6 100.00000
## 5      Finland        5         1  20.00000
## 6       France       18         5  27.77778
## 7      Iceland        2         2 100.00000
## 8       Israel        6         1  16.66667
## 9        Japan       10        10 100.00000
## 10   Lithuania       10         1  10.00000
## 11 New Zealand       14         2  14.28571
```

Also, we can see that, when analyzing the null values by country, there were some cases that have 100% of missing values in some attributes. This is why, we decided to drop **Costa Rica, Iceland and Japan** as they had attributes without values, and this would have not been useful for our project. In addition to them, we also decided to exclude Türkiye as we found poor the information contained in the Life.Satisfaction.(0-10) attribute.

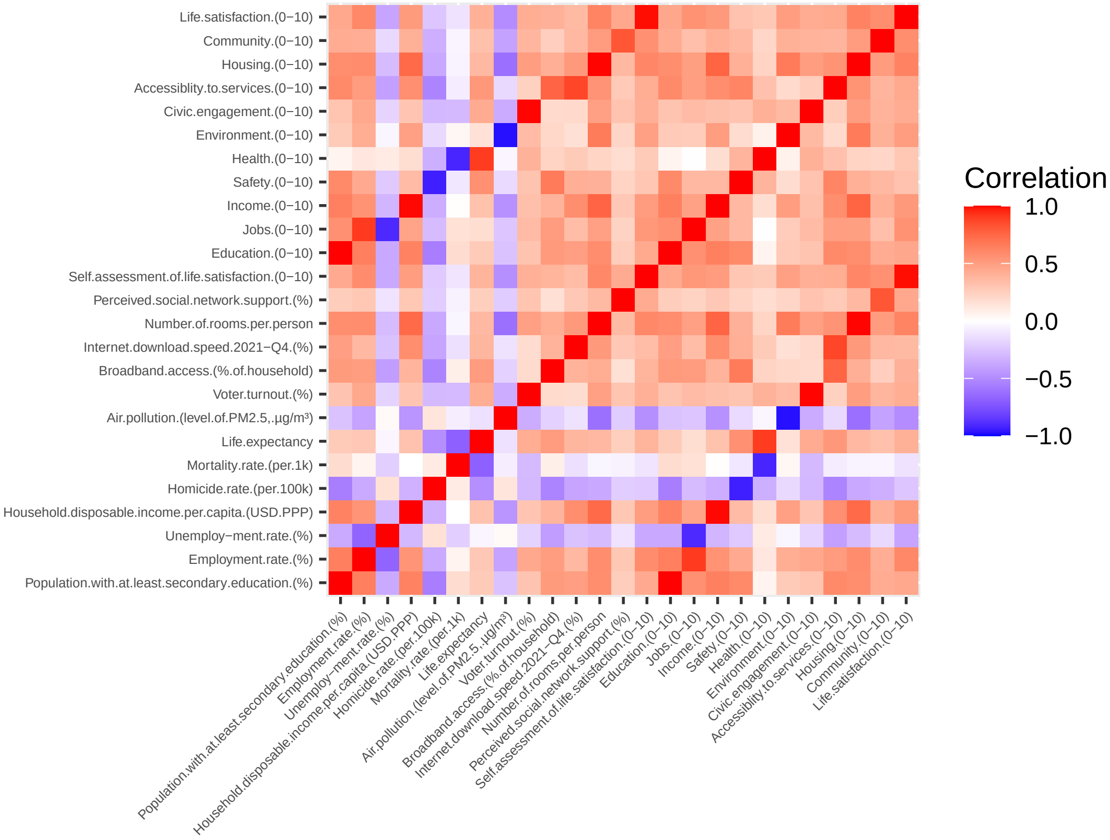
Moreover, for those cases with some null values, we decided to replace them with the **median of the country**. We decided this because the data is divided by city, so we could use the median of the rest cities of the country for those cities with null values. Furthermore, the median is a better option than the mean as we can avoid the influence of any possible outlier.

Finally, after all this data manipulation, we can see that now the data type of each attribute is correct.

```
##                                          Country
##                                      "character"
##                                           Region
##                                      "character"
##                                             Code
##                                      "character"
## Population.with.at.least.secondary.education.(%)
##                                        "numeric"
##                              Employment.rate.(%)
##                                        "numeric"
##                           Unemploy-ment.rate.(%)
##                                        "numeric"
## Household.disposable.income.per.capita.(USD.PPP)
##                                        "numeric"
##                         Homicide.rate.(per.100k)
##                                        "numeric"
##                          Mortality.rate.(per.1k)
##                                        "numeric"
##                                  Life.expectancy
##                                        "numeric"
##            Air.pollution.(level.of.PM2.5,.µg/m³)
##                                        "numeric"
##                                Voter.turnout.(%)
##                                        "numeric"
##                Broadband.access.(%.of.household)
##                                        "numeric"
##              Internet.download.speed.2021-Q4.(%)
##                                        "numeric"
##                       Number.of.rooms.per.person
##                                        "numeric"
##             Perceived.social.network.support.(%)
##                                        "numeric"
##      Self.assessment.of.life.satisfaction.(0-10)
##                                        "numeric"
##                                 Education.(0-10)
##                                        "numeric"
##                                      Jobs.(0-10)
##                                        "numeric"
##                                    Income.(0-10)
##                                        "numeric"
##                                    Safety.(0-10)
##                                        "numeric"
##                                    Health.(0-10)
##                                        "numeric"
##                               Environment.(0-10)
##                                        "numeric"
##                          Civic.engagement.(0-10)
##                                        "numeric"
##                  Accessiblity.to.services.(0-10)
##                                        "numeric"
##                                  Housing.(0-10)
##                                       "numeric"
##                                Community.(0-10)
##                                       "numeric"
##                        Life.satisfaction.(0-10)
##                                       "numeric"
```

Variables and Correlation

Continuing with the data manipulation, we decided to carry out a correlation analysis to determine the need of any further removal of attributes. For this, we decided to plot a correlation heatmap to rapidly see any pair of attributes highly correlated, and with thise, determine the removal of one of them.



As a result of this analysis, we decided to drop the following variables, because they can be explained by others (high correlation) and probably have less information than other correlated variables (this was checked manually with the dataset):

1. **Unemploy-ment.rate.(%):** -0.9 of correlation with Jobs.(0-10).
2. **Life.expectancy:** 0.9 of correlation with Health.(0-10).
3. **Internet.download.speed.2021-Q4.(%):** 0.87 of correlation with Accessiblity.to.services.(0-10).
4. **Perceived.social.network.support.(%):** 0.83 of correlation with Community.(0-10).
5. **Voter.turnout.(%):** 0.99 of correlation with Civic.engagement.(0-10).
6. **Air.pollution.(level.of.PM2.5,.µg/m³):** -0.97 of correlation with Environment.(0-10).
7. **Population.with.at.least.secondary.education.(%):** 0.99 of correlation with Education.(0-10).
8. **Household.disposable.income.per.capita.(USD.PPP):** 0.99 of correlation with Income.(0-10).
9. **Employment.rate.(%):** 0.92 of correlation with Jobs.(0-10).

Additionally, we decided to remove also Homicide.rate.(per.100k) and Mortality.rate.(per.1k) because we assumed that can be represented by Safety (0-10). Also we removed Broadband.access.(%.of.household) and Number.of.rooms.per.person as we did not consider it useful for the project. Last but not least, we removed Self.assessment.of.life.satisfaction.(0-10) as we directly used Life.satisfaction.(0-10).

Taking into account the resulting dataset, we decided to exclude those countries that have less than 10 cities in the dataset, as we have 10 variables plus the main variable (life satisfaction).

The last step of the data manipulation was to set the upper bound of the scale to 10, as there were some cases with decimals that ended up being a little bit over 10.

So the final dataset used in the models was the following, containing 295 tuples and 14 columns (10 are the attributes to analyze, plus the life satisfaction attribute):

```
## [1] 295  14
```

```
##    Country             Region              Code           Education.(0-10)
##  Length:295         Length:295         Length:295         Min.   : 0.009965
##  Class :character   Class :character   Class :character   1st Qu.: 4.313044
##  Mode  :character   Mode  :character   Mode  :character   Median : 7.330400
##                                                           Mean   : 6.497543
##                                                           3rd Qu.: 9.184019
##                                                           Max.   :10.000000
##   Jobs.(0-10)        Income.(0-10)       Safety.(0-10)      Health.(0-10)
##  Min.   : 0.008579   Min.   : 0.002448   Min.   : 0.00077   Min.   :0.009284
##  1st Qu.: 4.431959   1st Qu.: 0.916589   1st Qu.: 7.88044   1st Qu.:3.119418
##  Median : 6.861190   Median : 3.209448   Median : 9.45652   Median :5.956522
##  Mean   : 6.184882   Mean   : 3.486322   Mean   : 8.08465   Mean   :5.448041
##  3rd Qu.: 8.234613   3rd Qu.: 4.339201   3rd Qu.: 9.80978   3rd Qu.:7.647327
##  Max.   :10.000000   Max.   :10.000000   Max.   :10.00000   Max.   :9.504940
##  Environment.(0-10) Civic.engagement.(0-10) Accessiblity.to.services.(0-10)
##  Min.   : 0.00532   Min.   :0.009971        Min.   :0.008148
##  1st Qu.: 5.47264   1st Qu.:3.099481        1st Qu.:4.272850
##  Median : 7.31343   Median :4.960523        Median :6.630222
##  Mean   : 6.89282   Mean   :4.803227        Mean   :5.965565
##  3rd Qu.: 8.53234   3rd Qu.:7.078728        3rd Qu.:7.962655
##  Max.   :10.00000   Max.   :9.246560        Max.   :9.888185
##  Housing.(0-10)      Community.(0-10)   Life.satisfaction.(0-10)
##  Min.   : 0.008387   Min.   : 0.00998   Min.   : 0.009778
##  1st Qu.: 1.573034   1st Qu.: 4.41441   1st Qu.: 4.615385
##  Median : 4.494382   Median : 6.93694   Median : 6.538462
##  Mean   : 4.597146   Mean   : 6.28752   Mean   : 6.034132
##  3rd Qu.: 7.359551   3rd Qu.: 8.37838   3rd Qu.: 8.076923
##  Max.   :10.000000   Max.   :10.00000   Max.   :10.000000
```

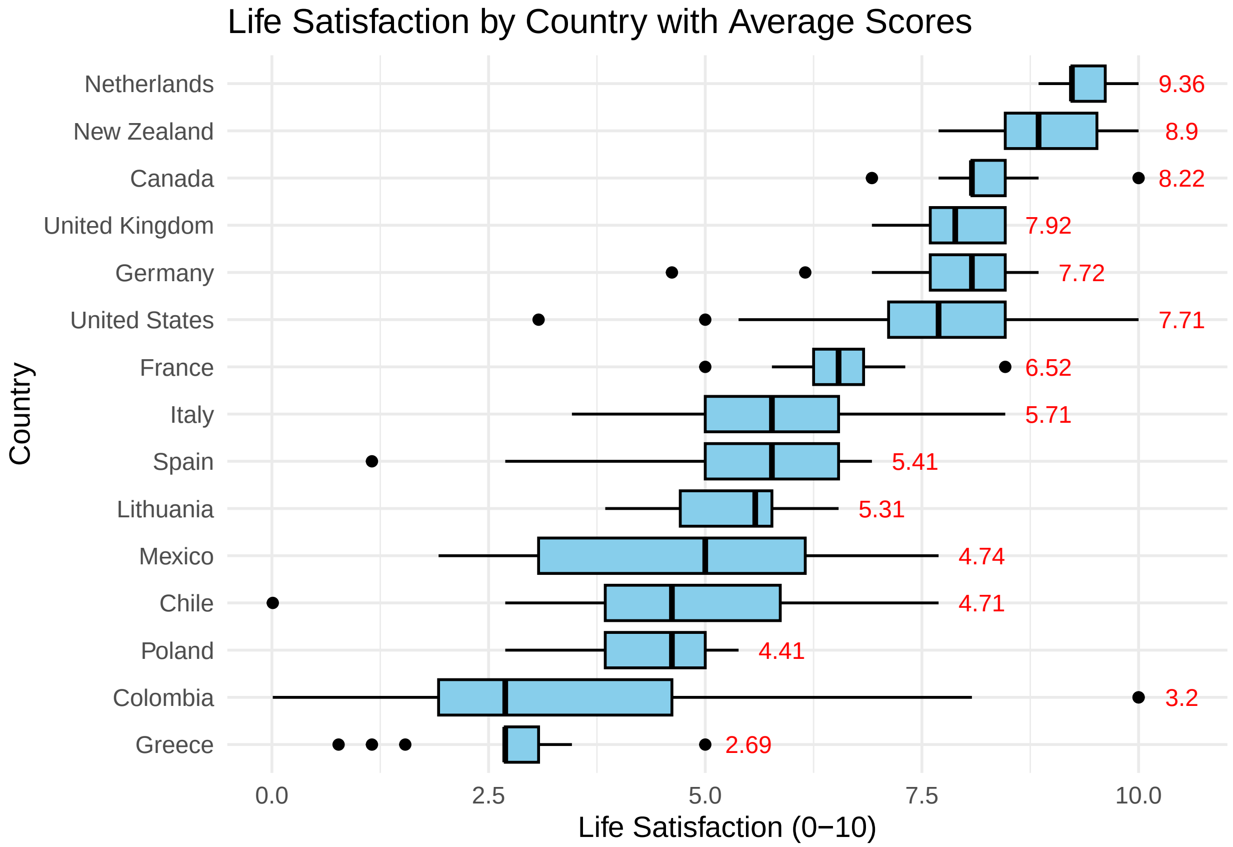
EDA

After completing a rigorous data cleaning process to remove redundancies and address multicollinearity, we continued with an Exploratory Data Analysis (EDA) to obtain insights about the factors influencing life satisfaction across 15 countries.

Life Satisfaction Distribution Across Countries

We began by analyzing how life satisfaction varies from one country to another. As illustrated in the box plot below, countries like the Netherlands, New Zealand, and Canada stood out with the highest average life satisfaction scores (9.36, 8.9, and 8.22, respectively). On the other hand, countries such as Greece, and Colombia reported the lowest scores.

Interestingly, in countries like Mexico and Colombia, we observed wide variability, suggesting significant regional differences within these nations. This led us to our **first key insight**: life satisfaction is not equally distributed across countries, and economic or social disparities likely play a role.



Understanding Relationships Between Metrics (Correlation Analysis)

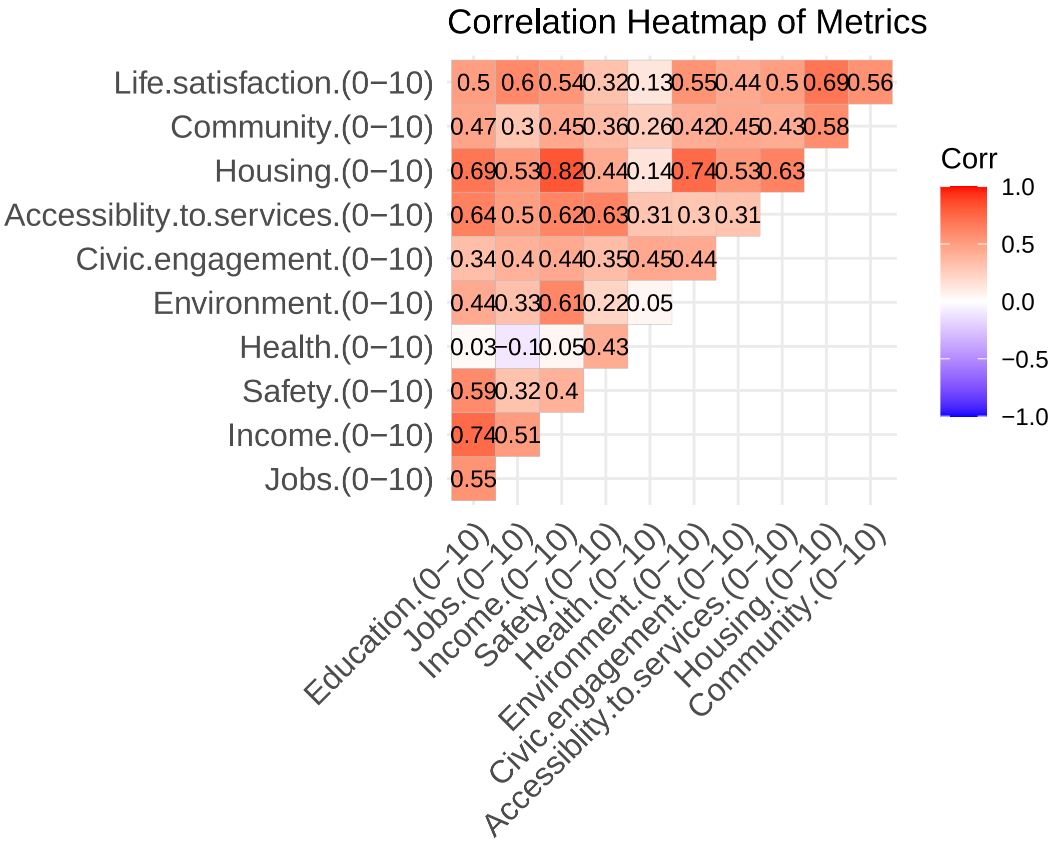
Deeper in our analysis, we explored the correlations between life satisfaction and other well-being metrics using a heatmap:

1. **Housing** had the strongest correlation with life satisfaction (0.70), showing that access to adequate housing and living conditions significantly influences well-being.
2. **Jobs** metric followed closely at 0.6, highlighting the importance of employment opportunities.

3. **Environemnt** and **Community** stood out with a correlation of 0.55 and 0.56 respectively, highlighting that a clean and sustainable environment and, connectedness and solidarity among groups in society are also an important part of life.

Surprisingly, **Income** showed a slightly lower correlation at 0.54, suggesting that while important, economic wealth is not the sole driver of life satisfaction.

This analysis revealed that housing conditions, employment opportunities, environment and community are more influential for life satisfaction than income alone.

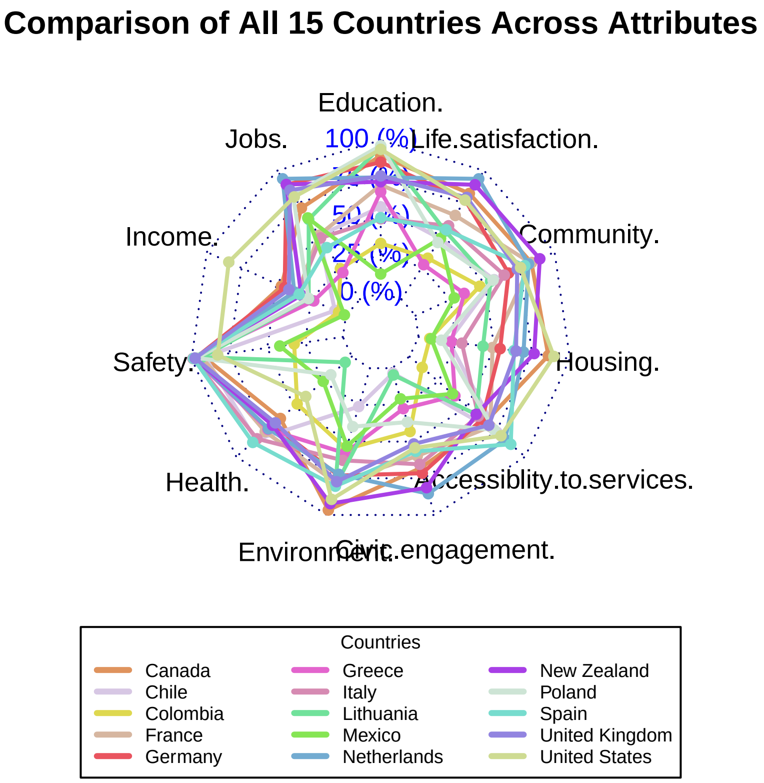


A Holistic Comparison of Countries (Radar Chart)

To compare the overall performance of all 15 countries, we created a radar chart that visualizes average scores across various attributes such as Environment, Jobs, Housing, and Education. Countries like the Netherlands and New Zealand demonstrated balance across multiple dimensions, outstanding in areas such as Education, Jobs, and Environment.

Conversely, countries like Greece struggled across several metrics, particularly in Income, Health, and Safety, which correlates with their lower life satisfaction scores.

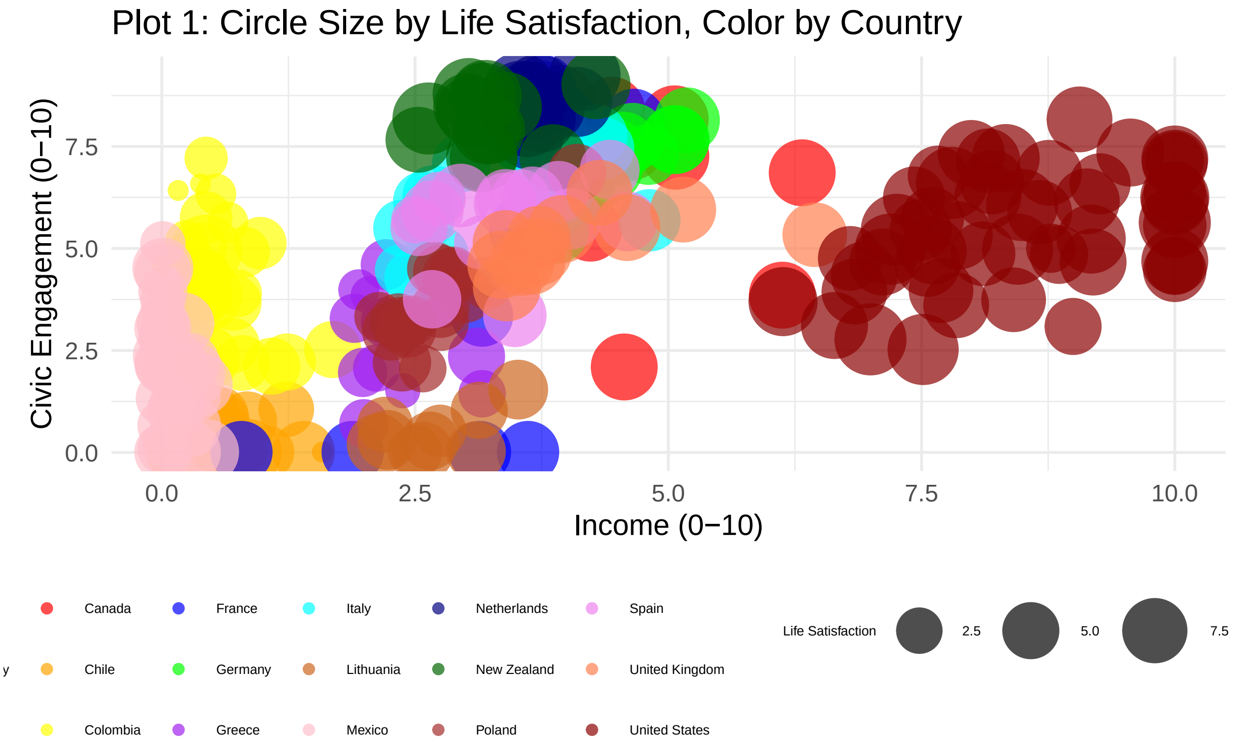
So from this visualization we can highlight some critical insight: Top-performing countries achieve balance across economic, environmental, and social dimensions, while lower-performing countries face challenges in multiple areas.



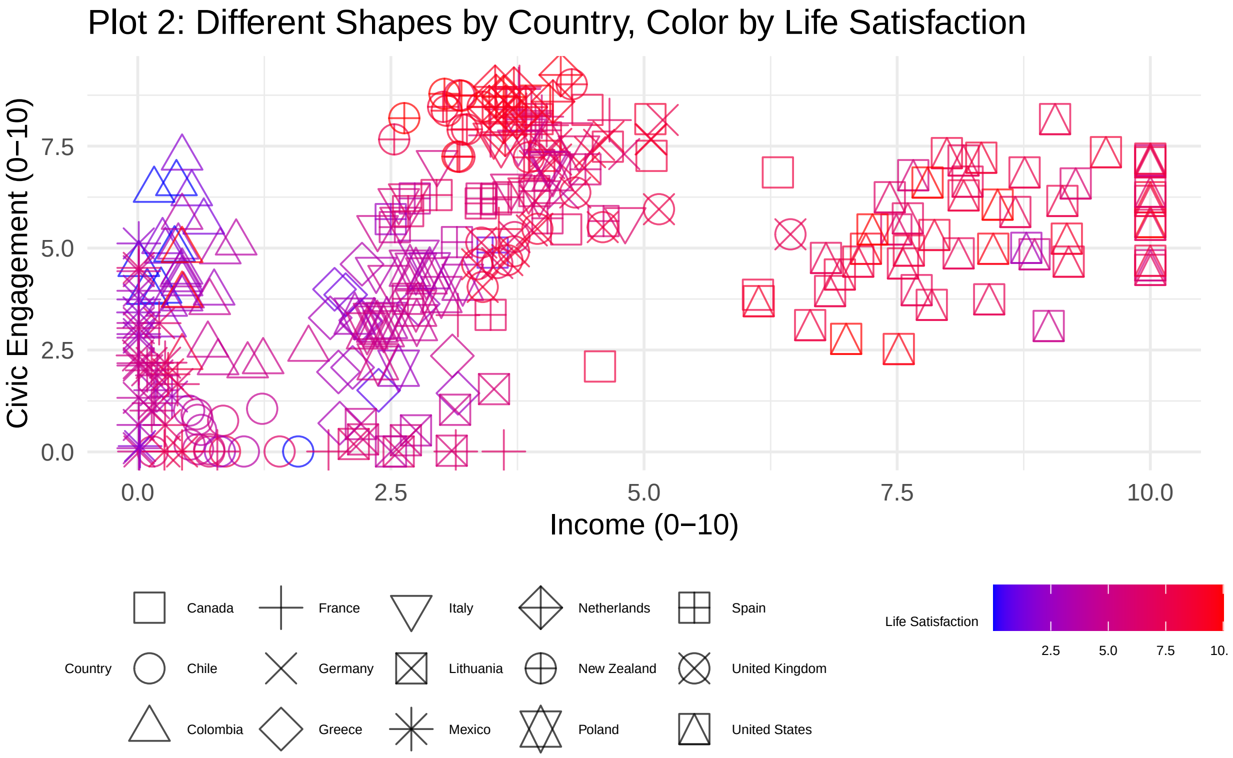
Visualizing Patterns Through Interactive Scatter Plots

Finally, we created two interactive plots to explore the relationships between Income, Civic Engagement, and Life Satisfaction:

1. **Plot 1: Circle Size by Life Satisfaction, Color by Country:** This plot revealed that countries with larger circles (higher life satisfaction) tend to have higher income and civic engagement levels. For instance, the Netherlands and the United States dominate the upper-right region of the plot. Conversely, smaller circles in the lower-left region highlight countries like Greece, where income and civic engagement remain low.



2. **Plot 2: Different Shapes by Country, Color by Life Satisfaction:** Here, each country is represented with a unique shape, and colors reflect their life satisfaction. Countries with higher satisfaction are visibly clustered in red, whereas lower-satisfaction countries are more dispersed, particularly toward the lower income range. These visualizations allowed us to see the clear relationships between income, civic engagement, and well-being, while also identifying disparities and regional patterns.


Plot 2: Different Shapes by Country, Color by Life Satisfaction
Civic Engagement (0−10)
Income (0−10)
Country
Canada
Chile
Colombia
France
Germany
Greece
Italy
Lithuania
Mexico
Netherlands
New Zealand
Poland
Spain
United Kingdom
United States
Life Satisfaction