CIP Project FS2025 - 202

Happiness around the World - 202

By Jade Bullock, Dóra Köhalmi, Ramona Kölliker

GitHub Repository: https://github.com/dkohalmi/CIP_FS25_202

1.0 Introduction

Happiness is a complex and multifaceted concept that is difficult to define, as it is influenced by both subjective and objective factors. Various elements, such as economic conditions, social support, physical and mental well-being, and personal freedom, all play a role in shaping happiness. It includes both how people feel (subjective well-being) and the conditions they live in (objective indicators).

This project examines various factors influencing happiness across different countries and over time. By integrating subjective happiness scores with objective country-level data, we have uncovered key insights into what contributes to overall well-being and have gained a deeper understanding of what people truly need to feel happy and fulfilled.

Our project aims to explore the key influences on happiness across countries by examining which factors matter most, how happiness differs between age groups, and the role that emotional experiences play in shaping overall well-being.

We focused on three main questions:

- What factors have the biggest impact on happiness?
- Is there a difference in happiness between old people and young people?
- Do people's experienced emotions affect happiness?

To answer these, we gathered data from trusted sources: the World Happiness Report, OECD Better Life Index, Gallup, and ILOSTAT. By combining data on how people feel with data on how they live, we aimed to get a fuller picture of global happiness.

2.0 Project Methodology

2.1 Data Sources and extraction

We collected data from several public sources, primarily using Python's Selenium package for automated web scraping and BeautifulSoup for HTML parsing. All scraping complied with website usage policies and no sensitive or personal data was extracted. Robots.txt was checked for all sites.

- **World Happiness Report 2024**

Data was scraped using the Selenium Python package from the World Happiness Data website. We automated dropdown interactions for all listed countries, capturing each country's overall happiness rank and average life evaluation score. Additionally, we extracted six explanatory factors (e.g., GDP, social support, freedom) from dynamically rendered tables using regular expressions and structured parsing. Waiting times were implemented to ensure full data loading. The dataset includes 164 countries and 20 variables.

- **Historical Happiness Scores (2012–2023)**

Data was downloaded directly from the Data Sharing | The World Happiness Report website as an xlsx file, which contained historical "Ladder score" life evaluation data for over 150 countries.

- **Happiness by Age (2021–2023)**

Data was scraped from Table 2.2 of the 2024 World Happiness Report. The script used XPath to locate the full table, then extracted life evaluation rankings for four age groups across 143 countries. Data includes rankings for the happiest and least happy age groups per country.

- **OECD Better Life Index**

Data was scraped from the OECD Better Life Index website using Python's Selenium package, which automated the process of accessing the Country section, extracting links from the dropdown menu, and visiting individual country pages. While the country pages follow a similar structure, some contain missing sections or inconsistent formatting. These issues required additional handling to accurately extract the relevant variables. The dataset includes 41 countries and 63 variables covering topics like housing, jobs, environment, and health.

- **Gallup Global Emotions:**

Data was scraped from Gallup's Global Emotions interactive dashboard. Selenium was used to click through each emotion toggle on the dashboard. Percentages of "Yes", "No", and "Don't Know" responses were collected per country using Selenium and parsed using BeautifulSoup. The dataset included over 142 countries.

- **Gallup Safety & Law and Order**

Data was downloaded directly from two publicly available Datawrapper CSV links embedded on Gallup's Law & Order page. These include perception-based safety and confidence-in-police scores. The dataset included over 142 countries.

- **ILOSTAT (Labor Indicators)**

A reusable scraping class was built to scrape data from ILOSTAT's iframe-embedded Datawrapper charts across seven labor topics: unemployment, employment, wages, working poverty, working time, labour productivity and safety. Selenium was used to switch into iframes, activate "Show more" buttons where available, and extract structured tables. For employment-specific scraping, a separate script handled dynamic HTML table structures using Firefox and BeautifulSoup due to the different page layout. The various datasets included between 70 -200 countries/regions. The Ilostat website was changed on 02/04/2025 after scraping to CSV. The raw CSV was used for further data processing.

2.2 Data Transformation and Cleaning

Each dataset underwent a tailored cleaning process to ensure consistency, accuracy, and usability for analysis. Across all datasets, we handled missing values, converted data types, standardized country names, and validated value ranges. We used the Pandas Python package for data manipulation tasks.

- **World Happiness Report 2024:**

Upon loading the raw CSV, placeholder values like "N/A" or dashes were treated as missing values. Seventeen countries lacked all data and were excluded from analysis. Two columns — "Healthy life expectancy rank" and "Healthy life expectancy value" — were completely empty and therefore dropped. All non-country object columns were cleaned and converted to floats. A new "Region" column was created by mapping each country to its region, based on the report's dashboard. Final cleaned data was saved and checked via descriptive statistics to ensure plausibility.

- **Historical Happiness Scores (2012–2023):**

This dataset was downloaded and required minimal cleaning. Ladder scores for the years 2020 to 2023 were merged with the 2024 dataset. Values were cross-checked to ensure consistency with the current-year report.

- **Happiness by Age (2021–2023):**

This dataset was relatively clean upon scraping. We verified column types and value ranges. No transformation was needed beyond standard formatting. Rankings for four age groups were retained, along with categorical indicators for the happiest and least happy groups.

- **Better Life Index:**

The collected data included numeric values in irregular formats (e.g., "~2.3 rooms" or "~76,235USD"), which were cleaned by removing non-numeric characters and standardizing the values. Missing data were handled during scraping, resulting in NaNs in the dataset. Most of these were in columns irrelevant to our analysis. Two missing values in the *Population* column (Costa Rica and Lithuania) were filled using official figures from the World Bank Data. After cleaning, the dataset was merged with the

Happiness Index data using a left join. Two country names were manually adjusted (*Slovak Republic* to *Slovakia*, and *Korea* to *Republic of Korea*) to ensure compatibility.

- **Gallup Safety & Emotions:**

For both datasets, raw values included percentages as strings (e.g., "85%"). These were cleaned and cast to numeric values. The Safety dataset was merged across two sources- personal safety and law & order scores, then checked for outliers (e.g. values > 100). Gallup Emotions data was pivoted from long to wide format using emotion-response pairs (e.g. joy_yes%), after removing the "Don't Know" category and standardizing labels.

- **ILOSTAT (7 Topics):**

Each of the seven labor datasets was scraped separately and cleaned using shared routines. Textual formatting (e.g. "$", "%", ",") was removed, and numeric fields were typecast. For indicators like injury rates or work hours, we validated values against expected thresholds. Outliers were flagged using IQR, and missing country values were reported. All datasets were eventually merged into one consolidated ILOSTAT dataset with aligned country names.

2.3 Data Analysis and Visualization

To investigate the factors influencing happiness, we applied a combination of exploratory data analysis (EDA), statistical modeling and clustering, supported by clear and intuitive visualisation. We used Python packages including Pandas and NumPy for data manipulation, Matplotlib, Seaborn, and Plotly for visualizations, Statsmodels and ScikitLearn for statistical modeling and clustering, and Streamlit to build an interactive web application. We created correlation heatmaps, regression and scatter plots, cluster maps, bar charts and boxplots to help us better understand the underlying relationships.

Exploratory and Comparative Analysis:

We began by summarizing and visually inspecting each dataset. Histograms, boxplots, and value counts helped us assess data quality and detect outliers. We used correlation matrices to explore relationships between happiness scores and predictors from the World Happiness Report, Gallup, ILOSTAT, and the OECD Better Life Index.

Statistical Modeling, Clustering, and Longitudinal Analysis:

We used Multiple Linear Regression models to assess how different factors influence happiness. Separate models were built to predict Happiness Index with the predictors:

- World Happiness Report variables (e.g. GDP, social support, freedom)
- Better Life Index topics (e.g. health, jobs, environment)
- Gallup emotions (e.g. enjoyment, learned, pain)
- ILOSTAT labor indicators (e.g. productivity, employment, minimum wages)

We also used VIF tests to check for multicollinearity and applied ANOVA and Tukey HSD tests to examine regional differences in factor contributions. To identify country groupings, we applied K-Means clustering on standardized features such as GDP and happiness scores.

A Streamlit app was also developed to provide interactive visualizations of the various datasets and key findings related to global happiness.

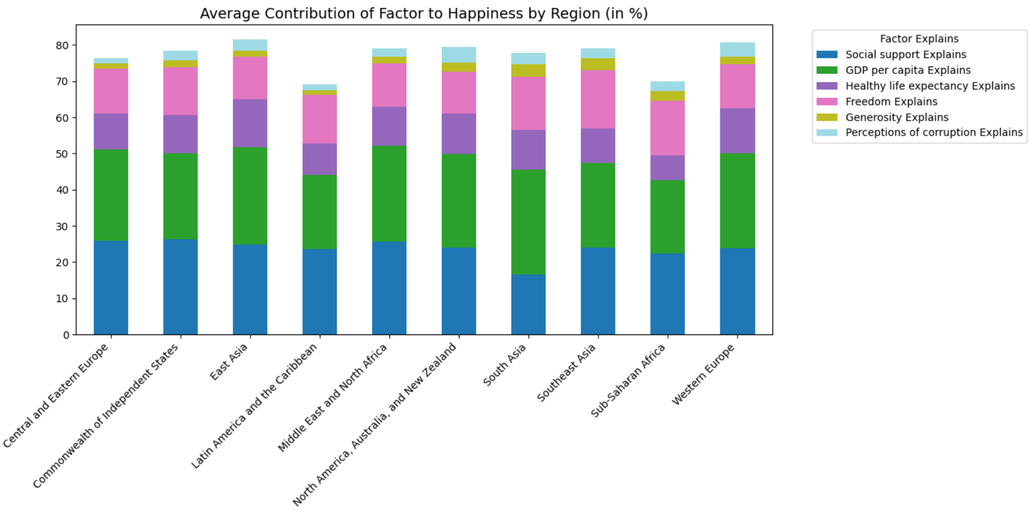
3.0 Analysis and Results

3.1 What Influences Happiness Most

Happiness Report Analysis:

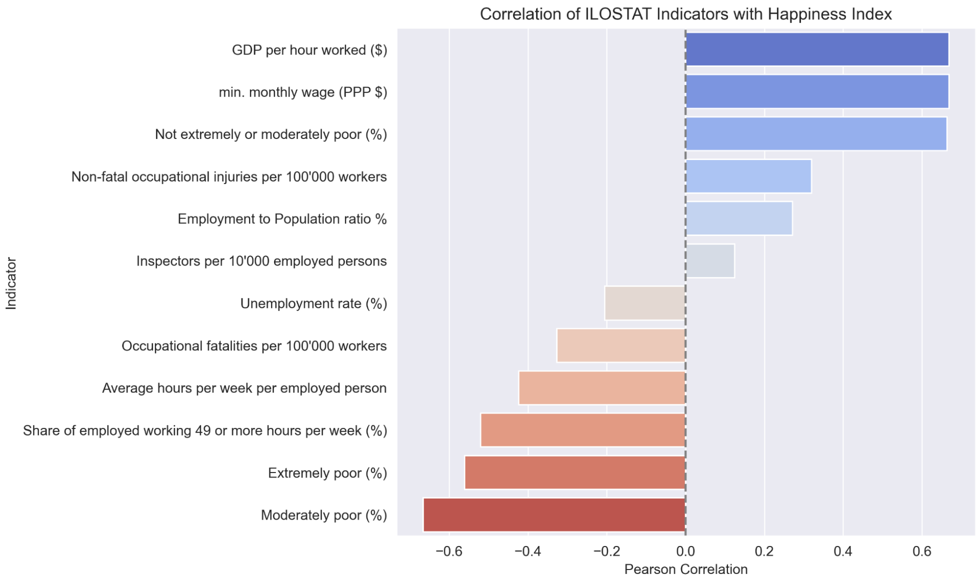
Social support and GDP per capita showed the strongest positive correlations with happiness. Freedom also had a meaningful positive association, while perceived corruption was negatively linked to happiness levels.

Linear regression analysis confirmed that GDP, social support, and freedom were statistically significant predictors of happiness scores. These variables collectively explained over 75% of the variation in life evaluation across countries.



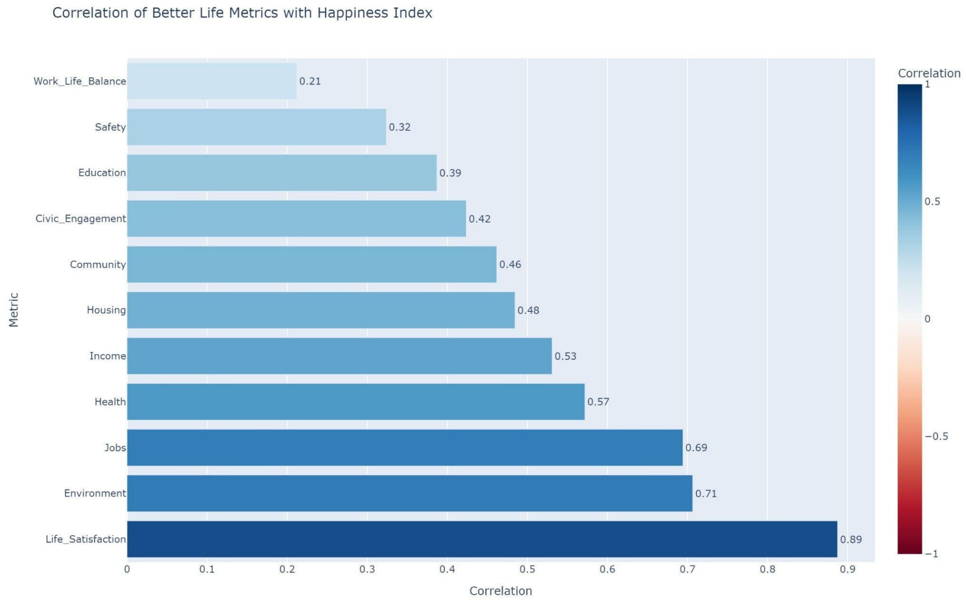
Labor Indicators from ILOSTAT:

Analysis of seven labor topics from ILOSTAT supported these findings by showing that objective work-related factors also contribute to well-being. Countries with higher labour productivity (GDP per hour worked) and employment-to-population ratios tended to have higher happiness scores. Conversely, high unemployment rates and working poverty were associated with lower life evaluation. While these indicators were not part of the World Happiness Report model, exploratory comparisons suggest that economic security, decent wages, and safe working conditions are important complements to subjective measures like freedom or social support.



Better Life Index Insights:

The OECD Better Life Index data revealed that Happiness Index is most strongly correlated with Environment, Jobs, and Health across the 41 countries analyzed. Income, Housing, and Community also showed notable positive correlations. Interestingly, Work-Life Balance had the weakest correlation with Happiness Index among the considered metrics. The findings of the linear regression model align with the insights gained from the correlation matrix, underscoring the significant influence of Jobs, Health, and Environment on Happiness levels within this dataset.



3.2 Regional Differences

Regional analysis indicated certain factors more heavily influenced happiness in certain regions. For example, health was more important in East Asia and North America, while freedom mattered more in South Asia. Contributions of social support and GDP were strong across all regions but varied slightly in magnitude.

Additional insights from the OECD Better Life Index data further highlighted how regional differences shape the drivers of happiness. In Western Europe and North America, Health and Environmental qualities were especially influential, aligning with higher investment in public services and environmental standards. In contrast, Jobs and Income were stronger predictors of happiness in Eastern Europe and Latin America, where economic stability remains a more central concern. These findings support the idea that while happiness is shaped by global factors, the relative importance of those factors differs by region, reflecting local contexts, priorities, and challenges.

3.3 Emotions and Age

Countries with higher levels of positive emotions, such as joy, gratitude, and interest, tended to report higher happiness scores. Negative emotions like anger and worry were more common in lower-ranked countries.

Age-based analysis from 2021 to 2023 showed that young people (under 30) were the happiest in most countries, while older adults (60+) often reported the lowest happiness levels. In several countries, the difference between age groups exceeded 30 ranking points, indicating substantial generational disparities. These patterns may reflect differences in health, opportunity, financial security, and social connection across life stages.

4.0 Conclusions

Our analysis shows that while income and economy matter, they aren't everything. Once a country reaches a certain income level, extra money doesn't lead to big gains in happiness. Instead, social support, emotional well-being, and personal freedom take center stage.

We also saw that emotions and safety perceptions strongly relate to happiness. Feeling safe and supported can often mean more to people than material wealth.

Age-based analysis revealed that younger people tend to report higher life satisfaction than older adults. In many countries, this generational gap is wide, sometimes exceeding 30 ranking points, highlighting important demographic dynamics in well-being. One contributing factor may be the increased experience of physical pain among older adults, as reflected in our Gallup Emotions data. Pain was one of the strongest emotional correlations of unhappiness, and generally its prevalence rises with age due to chronic health conditions or declining mobility. This connection highlights how both physical and emotional health contribute to the drop in happiness often observed later in life.

Importantly, the OECD Better Life Index data added a complementary perspective. Happiness scores were strongly correlated with Environment, Jobs, and Health- three key areas not directly covered in the WHR regression model. Income, Housing, and Community also showed notable positive correlations, whereas Work-Life Balance had a weak association. These findings suggest that economic stability alone isn't enough, people also need clean environments, decent work, and good health to thrive.

These findings reinforce the idea that happiness is not determined by any single factor but instead emerges from the complex interaction of emotional experiences, socioeconomic conditions, physical health, and environmental quality. By bringing together diverse datasets—from Gallup's emotional and safety metrics to ILO labor indicators and the OECD Better Life Index—we were able to explore happiness through multiple lenses. Importantly, the inclusion of both subjective perceptions and objective conditions revealed that neither emotional well-being nor economic prosperity alone can fully explain global variations in happiness. This multidimensional approach offers a more holistic understanding of what shapes people's lives and where policy efforts might most effectively be focused to support well-being across populations.

5.0 Limitations and Outlook

During data collection, we encountered several time-consuming issues, particularly due to changing website structures and inconsistent page layouts, which complicated the scraping process. Additionally, data coverage varied across sources, for example, the OECD dataset included only 41 countries, limiting the breadth of analysis. Analytical challenges included multicollinearity in regression models and cultural bias in emotion-based data, which may have affected cross-country comparability.

To enhance future research, we recommend expanding country coverage, conducting longitudinal analyses and integrating cultural indices to better interpret subjective and perception-based measures. Incorporating climate data and applying machine learning for predictive happiness modeling would also add value. Furthermore, country-specific case studies could help contextualize findings and reveal local nuances behind global trends.