Project Report: The Impact of Country Development on Happiness

Gence Genceli 34184

# 1. Motivation

I chose this project to understand what drives people’s happiness at a national level. By linking development indicators—such as health, education, and economic factors—to happiness scores, we aim to highlight which areas of investment can most effectively improve well-being globally.

# 2. Data Source

Our data come from multiple reputable sources for the period 2014–2024:  
- Human Development Index (HDI) and components (Life Expectancy, Education, GNI per Capita): UNDP Human Development Reports  
- Happiness Scores: World Happiness Report  
- Comfort Index (temperature deviation from 22 °C): World Bank climate data  
- Sports Success (5-year rolling FIFA ranking points): FIFA public rankings archive  
- Physical Activity Rates: WHO global health surveys  
- Internet Use Percentage: UNESCO ICT surveys  
- Income Inequality (Gini Index): World Bank World Development Indicators

# 3. Hypotheses

**Null Hypothesis H0:** Development of a country do not significantly influence happiness scores.

**H1: HDI vs. Happiness**

* *H₀:* There is no significant correlation between a country’s Human Development Index (HDI) and its average happiness score.
* *H₁:* There is a significant positive correlation between a country’s HDI and its average happiness score.

**H2: Comfort Index vs. Happiness**

* *H₀:* There is no significant correlation between a country’s Comfort Index (absolute deviation of average temperature from 22 °C) and its average happiness score.

*H₁:* There is a significant negative correlation between Comfort Index and happiness (i.e. countries whose climates are closer to 22 °C report higher happiness).

**H3: FIFA Success vs. Happiness**

* *H₀:* There is no significant correlation between a country’s total FIFA points and its average happiness score.
* *H₁:* There is a significant positive correlation between a country’s total FIFA points and its average happiness score.

**H4: Physical Activity vs. Happiness**

* *H₀:* There is no significant correlation between a country’s physical activity rate and its average happiness score.
* *H₁:* There is a significant positive correlation between physical activity rate and average happiness score.

**H5: Internet Use vs. Happiness**

* *H₀:* There is no significant correlation between the percentage of internet users and a country’s average happiness score.
* *H₁:* There is a significant positive correlation between internet‐use penetration and average happiness score.

**H6: Income Inequality vs. Happiness**

* *H₀:* There is no difference in mean happiness scores across countries grouped by income‐inequality category (Low, Medium, High Gini).
* *H₁:* There is a significant difference in mean happiness scores across income‐inequality categories (with more equal countries—Low Gini—having higher happiness)

# 4. Data Analysis

I followed a structured workflow comprising the following stages:

1. Cleaning & Merging: Standardized country names, filtered for 2014–2024, and handled missing values via interpolation or nearest-neighbor imputation.

Exploratory Data Analysis: Generated histograms, boxplots, and scatter plots to visualize distributions and pairwise relationships.

1. Computed summary statistics (mean, median, standard deviation) for each variable.
2. Generated histograms for all continuous predictors and the happiness score to assess distribution shapes.
3. Created boxplots to compare key variables across income-inequality categories.
4. Plotted scatter plots with trendlines for pairwise relationships (e.g., HDI vs. Happiness, Comfort Index vs. Happiness).
5. Examined outliers and data quality issues visually and through interquartile range rules.
6. **Correlation Assessment**: Computed Pearson correlation matrix and plotted a heatmap to detect multicollinearity.
7. **Statistical Testing & Modelling:** Conducted ANOVA for income-inequality categories and fitted multiple regression models (Linear, Ridge, Random Forest), evaluating each with RMSE, MAE, and R².

# 5. Findings

* Strong positive correlation between HDI and happiness (r ≈ 0.80).
* Life expectancy and education have the largest individual impacts.
* Climate comfort shows a moderate positive relationship.
* Internet use and physical activity rates both correlates positively with happiness.
* Lower income inequality (low Gini) associates with higher median happiness.

# 6. Graphs & Visualizations and Their Explanations

A graph of happiness score

AI-generated content may be incorrect.

[Figure 1: Actual vs. Predicted (Random Forest)]

This scatter plot compares actual happiness scores (x-axis) with the scores predicted by the Random Forest regression model (y-axis). The dashed diagonal line represents perfect predictions. Most points lie close to the line, indicating that the Random Forest model captures the relationship very well, achieving the lowest overall prediction error among the models tested.

A graph with blue dots

AI-generated content may be incorrect. [Figure 2: Actual vs. Predicted (Ridge Regression)]

Here, the scatter plot shows predictions from the Ridge regression model against actual happiness scores. While the points generally follow the diagonal trend, they are more dispersed compared to the Random Forest plot, revealing slightly higher prediction errors due to the model's trade-off between bias and variance.

A graph with blue dots

AI-generated content may be incorrect.

[Figure 3: Actual vs. Predicted (Linear Regression)]

This plot illustrates the performance of the simple linear regression model. The greater spread of points around the diagonal line indicates higher prediction errors and a lower R² value than the other models, reflecting the limitations of a purely linear approach for this dataset.

A colorful squares with white text

AI-generated content may be incorrect.[Figure 4: Correlation Heatmap of Key Variables]

The heatmap displays Pearson correlation coefficients among our main variables. Key observations include a strong positive correlation between HDI and happiness (r ≈ 0.80), a moderate positive correlation for internet use (r ≈ 0.66), and a moderate negative correlation for physical activity rate (r ≈ -0.44). Income inequality (Gini index) shows a weak negative correlation (r ≈ -0.21), indicating more unequal countries tend to report slightly lower happiness.

A graph of happiness score

AI-generated content may be incorrect.

[Figure 5: Histogram of Happiness Scores]

The histogram of happiness scores reveals a distribution concentrated between 4 and 7, with a noticeable peak around 6. This suggests that most countries report moderate-to-high happiness levels, with fewer extremes at the lowest and highest ends.

A diagram of a graph

AI-generated content may be incorrect.

[Figure 6: Boxplot – Income Inequality vs. Happiness]

This boxplot compares happiness distributions across three income-inequality categories: Low (Gini < 30), Medium (Gini 30–40), and High (Gini > 40). Countries with low income inequality have the highest median happiness, while those with high inequality show lower median scores and greater variability, highlighting the social impact of income distribution.

# 7. Validity of the Hypotheses

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hypothesis** | **Test** | **Statistic** | **p-value** | **Decision** |
| **H1: FIFA Success vs. Happiness** | Pearson correlation | r = 0.30 | 2.13 × 10⁻²⁷ | Reject H₀ (significant positive link) |
| **H2: HDI vs. Happiness** | Pearson correlation | r = 0.80 | 6.24 × 10⁻³⁰⁰ | Reject H₀ (strong positive link) |
| **H3: Comfort Index vs. Happiness** | Pearson correlation | r = 0.43 | 1.60 × 10⁻⁶¹ | Reject H₀ (moderate negative link) |
| **H4: Physical Activity vs. Happiness** | Pearson correlation | r = –0.44 | 1.37 × 10⁻⁵⁵ | Reject H₀ (significant negative link) |
| **H5: Internet Use vs. Happiness** | Pearson correlation | r = 0.66 | 1.13 × 10⁻¹⁴ | Reject H₀ (strong positive link) |
| **H6: Income Inequality vs. Happiness** | One-way ANOVA | F = 22.68 | 4.65 × 10⁻¹⁰ | Reject H₀ (groups differ significantly) |

# 8. Machine Learning Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **CV R²** | **RMSE** | **MAE** | **Test R²** |
| Linear Regression | 0.657 ± 0.042 | 0.663 | 0.504 | 0.672 |
| Ridge Regression | 0.657 ± 0.042 | 0.662 | 0.504 | 0.672 |
| Random Forest Regression | 0.898 | 0.343 | 0.233 | 0.912 |

* **Best Model:**  
  Random Forest Regression (highest Test R² of 0.912 and lowest RMSE of 0.343)

# 9. Limitations & Future Work

**Limitations:**  
- Some years and countries lacked full survey data, requiring interpolation that may introduce bias.  
- Regression analyses show association, not causation; unmeasured confounders (e.g., cultural factors) may influence results.

**Future Plans:**  
-To deal with missing data, use sensitivity analyses and sophisticated imputation techniques.   
-Create periods of time explanations to investigate changes within a nation.   
-Include qualitative factors from other surveys, such as freedom and trust.   
-Create a dashboard with interactive elements to investigate the factors that influence happiness at the national level.

# 10. Software and Libraries

* pandas: Data manipulation and analysis
* numpy: Numerical computing and array operations
* matplotlib: Data visualization
* scipy: Statistical testing and scientific computing
* scikit-learn: Machine learning modelling and evaluation.

**11. References**

**LLM Assistants Consulted**

* ChatGPT
* Claude

While debugging codes, and rephrasing some words for this report 😊

**Datasets and Course Materials**

* Processed\_2014\_2024.csv
* DSA210 Lecture Slides (w1-w7), lecture codes and Recitation Materials