DSA-210 FINAL REPORT

ZEYNEP OKKIRAN

32304

Gender Life Expectancy Gap and Different Factors on Happiness Levels

SUPERVISED BY SELİM BALCISOY

SABANCI UNIVERSITY

Introduction

The principal objective of this project is to examine the relationship between gender-based disparities in life expectancy and national happiness scores. Additionally, this study aims to assess whether other social factors—namely economic performance, freedom, and generosity—also significantly influence happiness scores beyond the life expectancy gap.

What did I do?

The project began with comprehensive exploratory data analysis, followed by rigorous data cleaning to ensure the integrity and consistency of the dataset. Next, a suite of advanced visualization techniques was applied to reveal distributional patterns, trends, and potential outliers. A Pearson correlation test was then conducted as part of hypothesis testing to quantify the strength and direction of relationships between key variables. Finally, several regression-based machine learning models were developed and evaluated—using metrics like RMSE and R^2, and validated through cross-validation—to predict national happiness scores based on both life-expectancy differences and additional socio-economic factors.

Hypothesis

- Null Hypothesis(H0): There is no significant correlation between the difference in life expectancy between women and men and the happiness score of countries.
- Alternative Hypothesis(H1): Countries where women live significantly longer than men tend to have higher happiness levels.

Parameters in the Report

 Country: Name of the nation.

 Happiness Rank: Position in the global happiness ranking (1 = happiest).

 Happiness Score: Self-reported well-being on a standardized scale.

 Economy (GDP per Capita): The GDP-per-person component's contribution to the overall happiness score.

 Health (Life Expectancy): The life-expectancy component's contribution to the overall happiness score.

 Freedom: Index of perceived freedom to make life choices.

 Generosity: Measure of charitable giving in the country.

 Female Life Expectancy: Average years a newborn female is expected to live.

 Male Life Expectancy: Average years a newborn male is expected to live.

 Overall Life Expectancy: Average years a newborn (both sexes) is expected to live.

 Life Expectancy Difference: Female minus male life expectancy (years).

Data visualizations

Figure 1 presents a preliminary scatter plot of national Happiness Score versus average life expectancy, revealing a clear positive association: countries with higher self-reported well-being generally enjoy longer lifespans. While most points cluster between 70–83 years of expectancy and 5.0–7.0 happiness, some outliers (e.g., moderate happiness at low expectancy or vice versa) hint that additional socio-economic variables influence well-being. This initial view supports a formal correlation analysis and motivates our subsequent multivariate modeling.

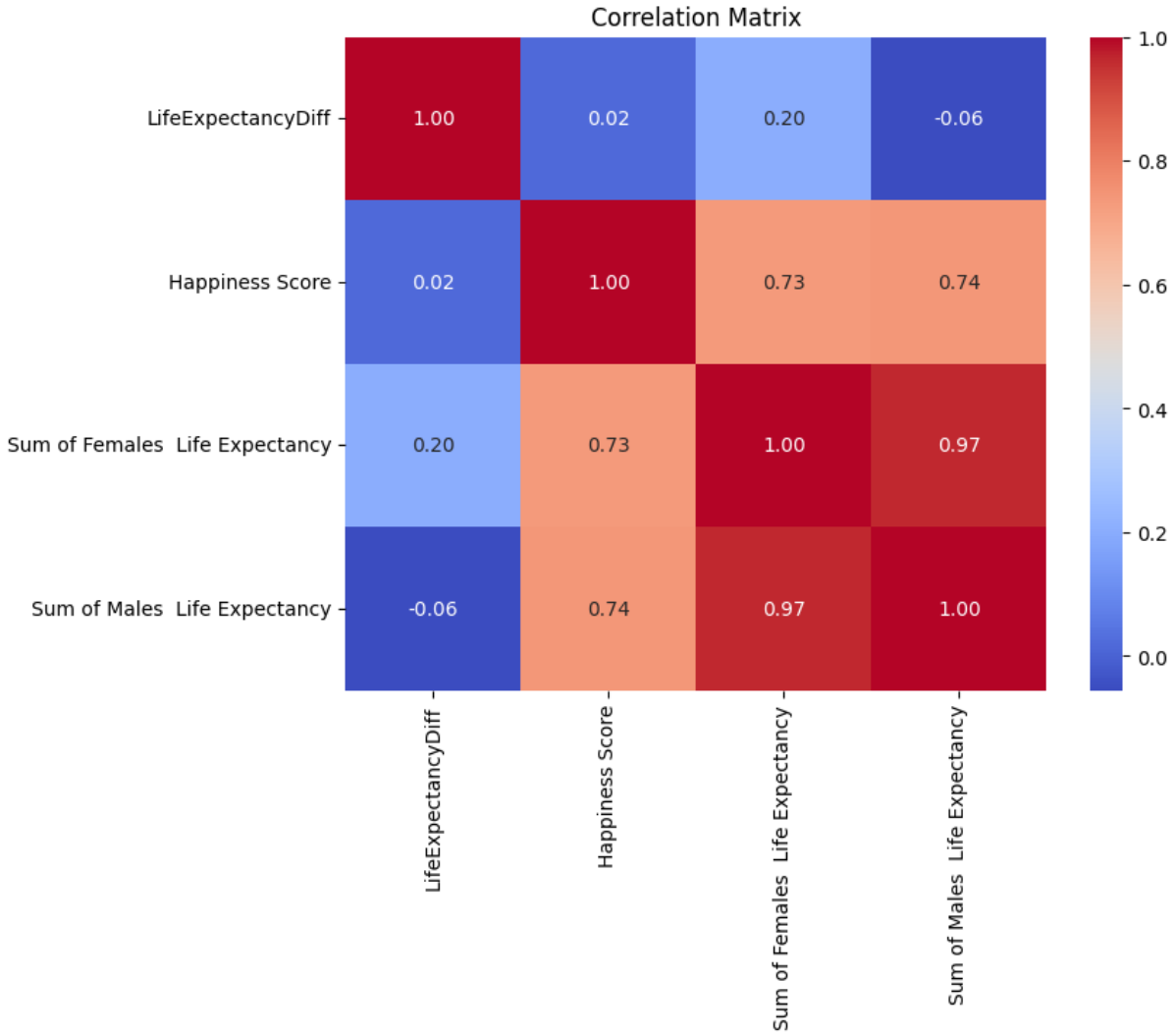


Figure 2 displays a heatmap of the pairwise values among our four life-expectancy metrics (with each variable normalized so that self-comparisons sit at 1.00 on the diagonal). Key off-diagonal figures include:

- **Female vs. Male Life Expectancy:** 0.97 (deepest red), the highest non-diagonal value.
- **LifeExpectancyDiff vs. Male Life Expectancy:** –0.06 (deepest blue), the lowest value.
- **LifeExpectancyDiff vs. Female Life Expectancy:** 0.20 (light blue).
- **LifeExpectancyDiff vs. Happiness Score:** 0.02 (very pale blue), indicating a technically slight **positive** association—but effectively negligible.
- **Happiness Score vs. Female Life Expectancy:** 0.73 (orange-red).
- **Happiness Score vs. Male Life Expectancy:** 0.74 (orange-red, almost identical).

This gradient—from –0.06 through 0.02/0.20 up to 0.73/0.74 and peaking at 0.97—highlights which metric pairs share similar magnitudes and which stand in stark contrast; notably, the minimal link between the gender gap in life expectancy and happiness suggests that absolute longevity, rather than its sex-based disparity, drives the observed well-being relationship.

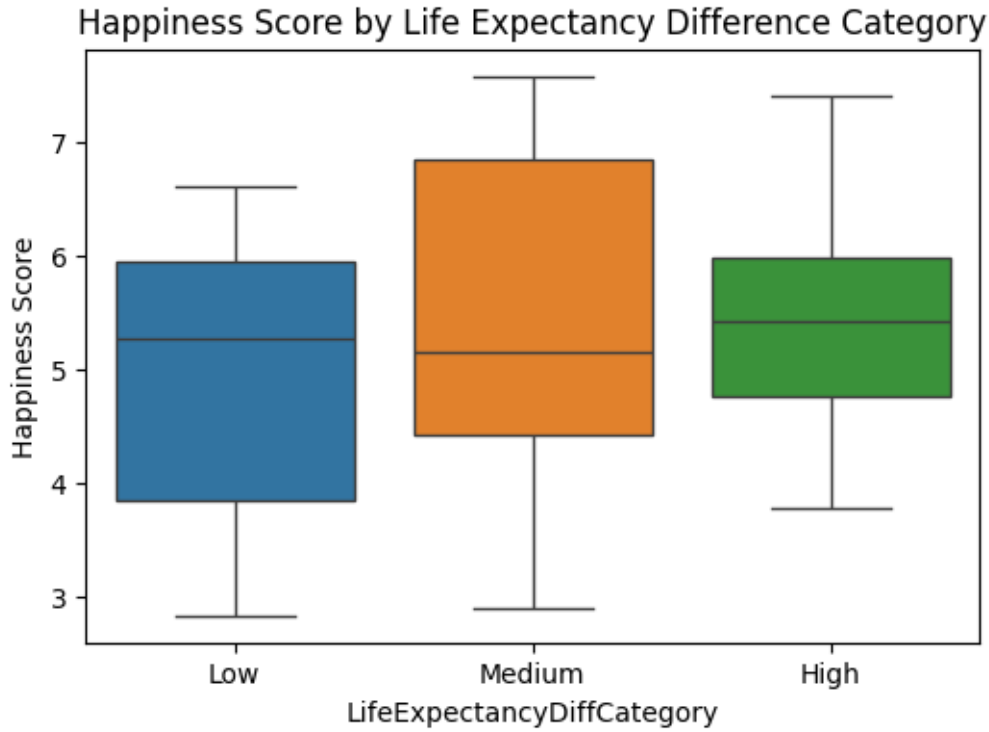


Figure 3 compares the distribution of national Happiness Scores across three Life Expectancy Difference categories (Low, Medium, High). All three boxes overlap substantially, with medians clustered around 5.2–5.4 points (Low ≈5.3, Medium ≈5.2, High ≈5.4). The Medium-gap group shows the greatest variability (range ≈2.9–7.6), while the Low and High groups span roughly 2.8–6.6 and 3.8–7.3, respectively. Despite a slightly higher median in the High category, the interquartile ranges overlap heavily and the differences in central tendency are under 0.3 points, indicating that the gender gap in life expectancy exerts minimal practical influence on overall happiness.

Hypothesis Testing

I chose Pearson's correlation test because my primary interest was in assessing whether there exists a linear relationship between two continuous variables—namely, the gender gap in life expectancy (LifeExpectancyDiff) and national Happiness Scores. Pearson's r is the standard metric when both variables are measured on an interval scale and are approximately normally distributed, as is the case with country-level life expectancy and happiness data. Before applying the test, I visually inspected scatter plots and distributional histograms to confirm that neither variable displayed extreme skewness or heavy outliers that would violate Pearson's assumptions. By focusing on Pearson's r, I directly quantified both the strength and direction of any linear association, enabling straightforward hypothesis testing against the null hypothesis of zero correlation.

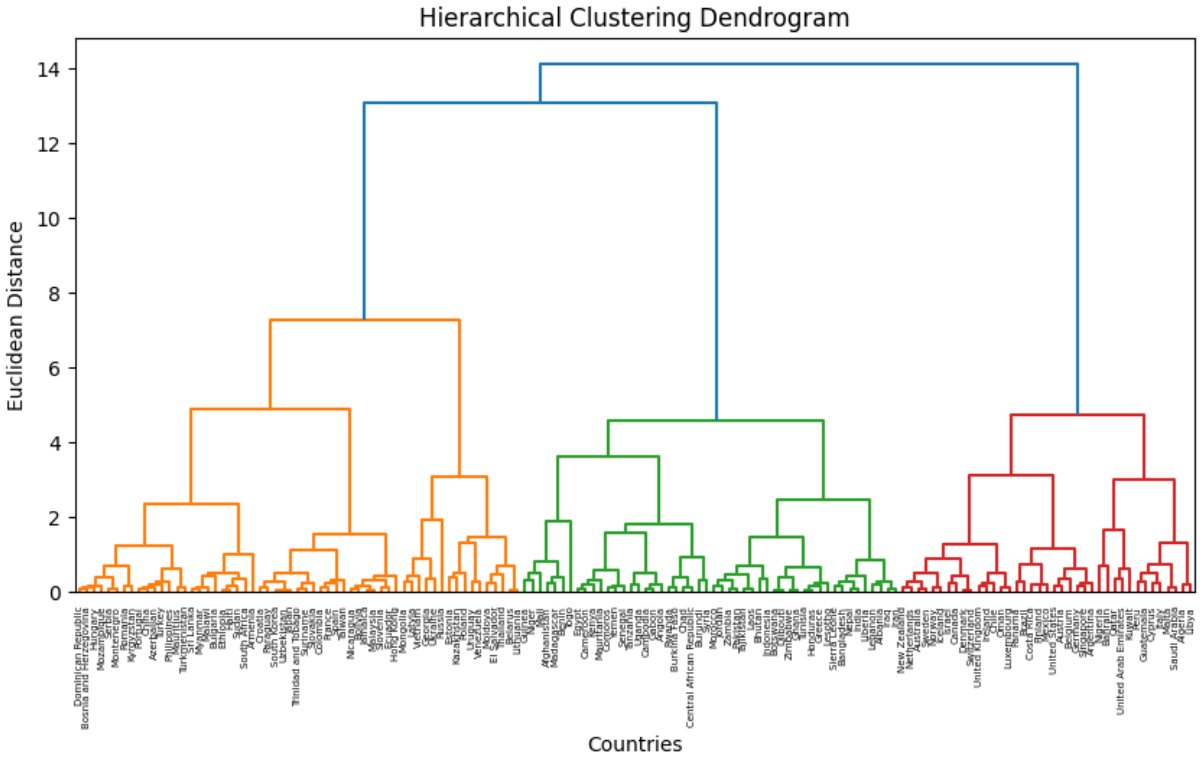
Results of hypothesis testing

The analysis yielded a correlation coefficient of $r=0.0172$ and a p-value of 0.8358. Because r is essentially zero and the p-value greatly exceeds the conventional $\alpha=0.05$ threshold, we **fail to reject the null hypothesis**. In practical terms, this means there is **no statistically significant** linear association between the magnitude of the life-expectancy gap and a country's happiness score. Consequently, the gender disparity in longevity appears to have negligible influence on national well-being, reinforcing our decision to concentrate subsequent modeling efforts on absolute life expectancy and other socio-economic predictors.

Unsupervised Model: Hierarchical Clustering

In the unsupervised learning phase of this individual project, I applied **hierarchical clustering** to the socio-economic and life-expectancy variables (LifeExpectancyDiff, Happiness Score, Economy, Health, Freedom, Generosity) in order to reveal natural groupings among the ~150 countries without predefining a cluster count. Hierarchical clustering was chosen because it handles continuous variables on different scales, does not require specifying k in advance, and produces an intuitive dendrogram for visual interpretation. As shown in the attached dendrogram, cutting the tree at a Euclidean distance of roughly 7 yields three coherent clusters: (1) very high-income, high-happiness nations; (2) large developed and resource-oriented economies; and (3) emerging-market and lower-income countries. This clustering outcome corroborates earlier analyses by demonstrating that absolute life expectancy and

broader economic and social factors drive country similarity, whereas the gender gap in longevity has minimal influence on these clusters.



After using the dendrogram to identify that three clusters best reflect the underlying structure, I turned to **AgglomerativeClustering** with n_clusters = 3 to convert those visual groupings into concrete labels. Unlike the purely visual dendrogram, AgglomerativeClustering applies the same bottom-up merging logic in code—ensuring consistency with the hierarchy—while yielding reproducible numeric cluster IDs. This step enabled precise quantification of cluster characteristics (e.g., group sizes and average Happiness Scores), facilitating objective comparison across the three segments.The resulting cluster assignments appear below (showing the first ten rows), along with cluster sizes and mean Happiness Scores:

Country	Hierarchical_Cluster
Switzerland	1
Iceland	1
Denmark	1
Norway	1
Canada	1
Finland	1
Netherlands	1
Sweden	1
New Zealand	1
Australia	1

Cluster sizes

- Cluster 0: 59 countries
- Cluster 1: 39 countries
- Cluster 2: 50 countries

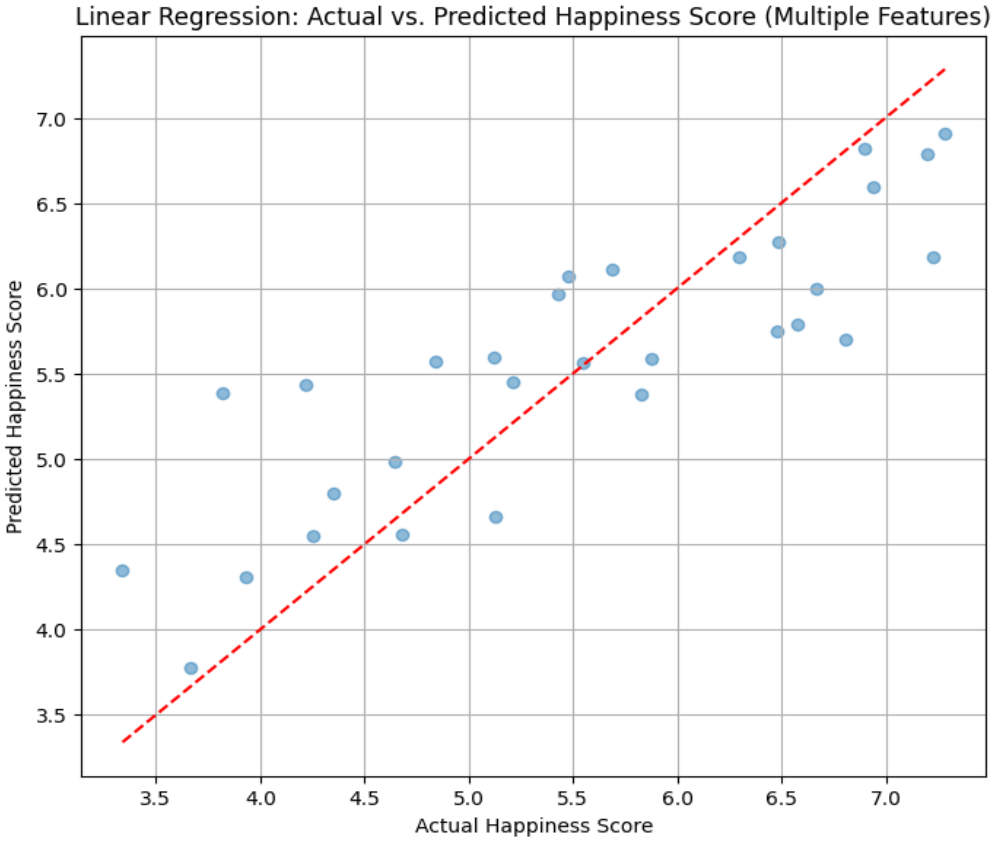
Average Happiness Score by cluster

- Cluster 0: 5.4509
- Cluster 1: 6.7730
- Cluster 2: 4.2803

Cluster 1 clearly corresponds to the highest-income, highest-happiness nations (e.g. Switzerland, Iceland, Denmark…), Cluster 0 aligns with large developed and resource-oriented economies (intermediate happiness), and Cluster 2 groups emerging-market and lower-income countries (lowest happiness). These flat clusters therefore validate the three-group structure visualized in the dendrogram and reinforce that absolute socio-economic and health factors—rather than the gender life-expectancy gap—drive the main country-level distinctions.

Machine learning: Regression models

- Linear Regression



Linear Regression MSE: 0.4028371292165504

Linear Regression R^2: 0.69933675117131

Linear Regression Coefficients:

LifeExpectancyDiff: 0.019319348326146176

Economy (GDP per Capita): 1.5087548076113928

Health (Life Expectancy): 0.7525620847497514

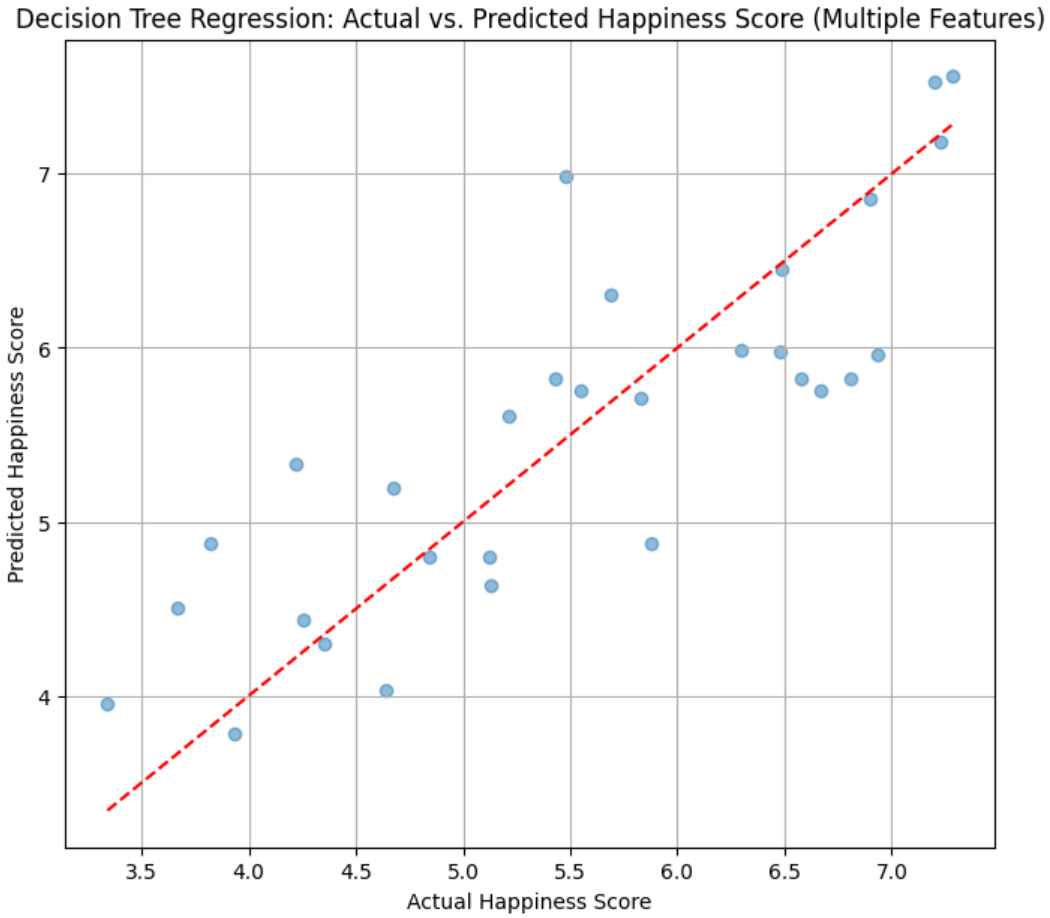
Freedom: 2.214159212510876

Generosity: 0.5220772867000429

Intercept: 2.4541017241707106

The regression coefficients rank as follows: Freedom (2.21) > Economy (1.51) > Health (0.75) > Generosity (0.52) > Life Expectancy Difference (0.02), indicating that **freedom is by far the strongest predictor, while the gender gap in longevity has a negligible effect on happiness**. Combined with an **MSE of 0.40** (RMSE ≈ 0.63) and an **R^2 of 0.70**, the model explains 70% of the variance in national happiness and delivers predictions that are, on average, within about 0.6 points of the true scores, indicating a good model fit.

- Decision Tree Regression

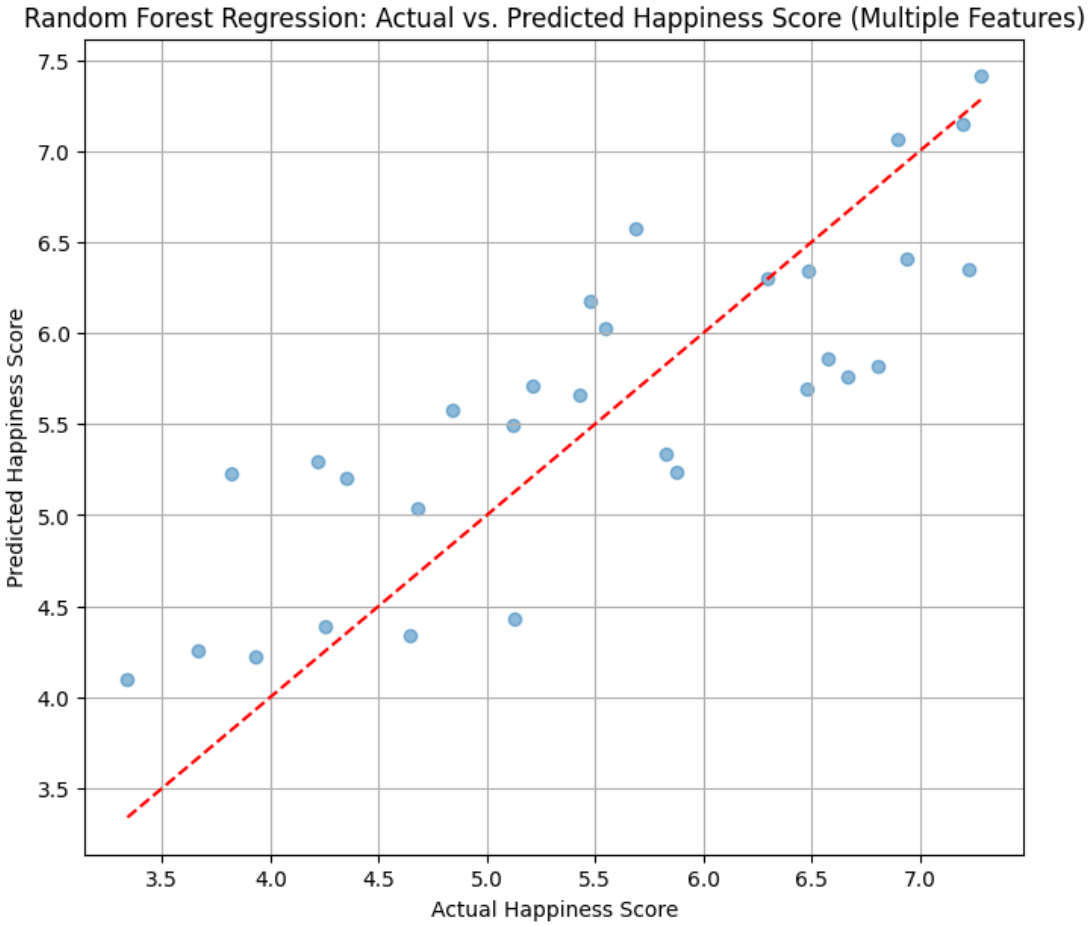


Decision Tree Regression MSE: 0.4123784666666667

Decision Tree Regression R^2: 0.692215437598497

The Decision Tree Regression achieves an **MSE of 0.412** (RMSE ≈ 0.64) and an **R^2 of 0.692**, meaning it explains about 69.2% of the variance in national happiness scores. On average, its predictions are off by roughly 0.64 points—less accurate than the linear model.

- Random Forest Regression



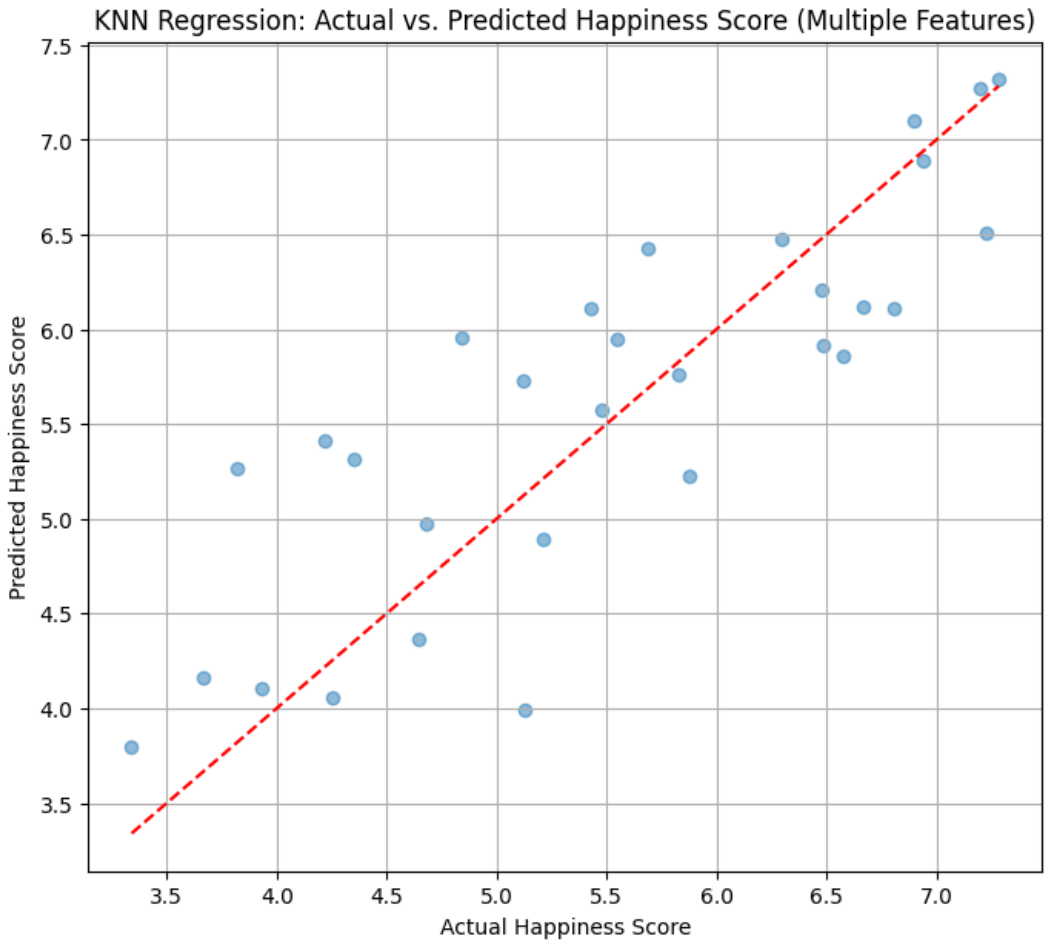
Random Forest Regression MSE: 0.4270671978999994

Random Forest Regression R^2: 0.6812522931078828

The Random Forest Regression yields an **MSE of 0.4271** (RMSE ≈ 0.65) and an **R^2 of 0.6813**, meaning it explains about 68.1% of the variance in national happiness scores. On average, its predictions deviate by roughly 0.65 points

indicating a solid but slightly weaker performance compared to the linear and decision tree models.

- KNN Regression



KNN Regression MSE: 0.4020466640000001

KNN Regression R^2: 0.6999267261831872

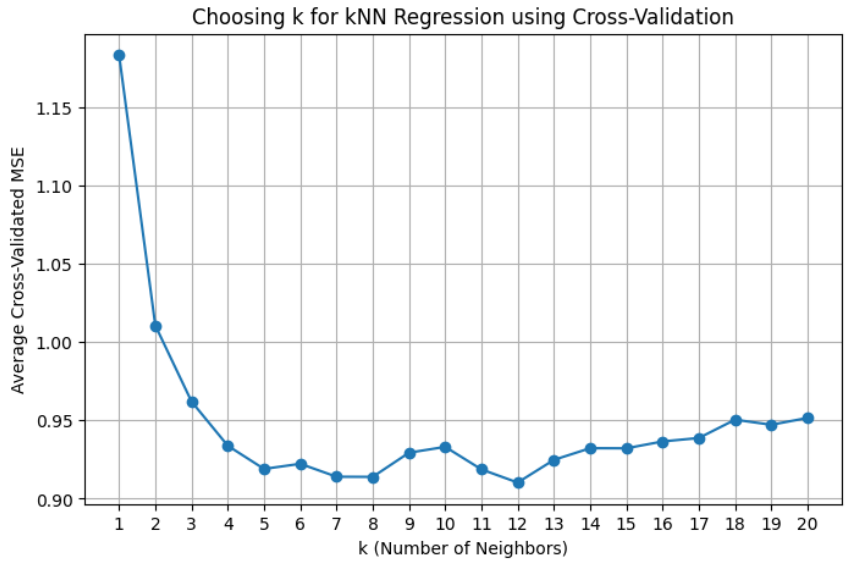
The KNN Regression achieves an **MSE of 0.4020** (RMSE ≈ 0.63) and an **R^2 of 0.6999**, meaning it explains about 70% of the variance in national happiness. With the lowest error and highest R^2 among the models tested, it slightly outperforms the others in predictive accuracy.

Model	Mean Squared Error(MSE)	R^2
Linear Regression	0.4028	0.6993
Decision Tree	0.412	0.692
Random Forest	0.427	0.681
KNN	0.4020	0.6999

This table shows that KNN Regression, with the lowest MSE and highest explained variance (R^2), outperforms the other models and delivers the most accurate predictions.

Model Improvement: KNN Regression with Optimal k

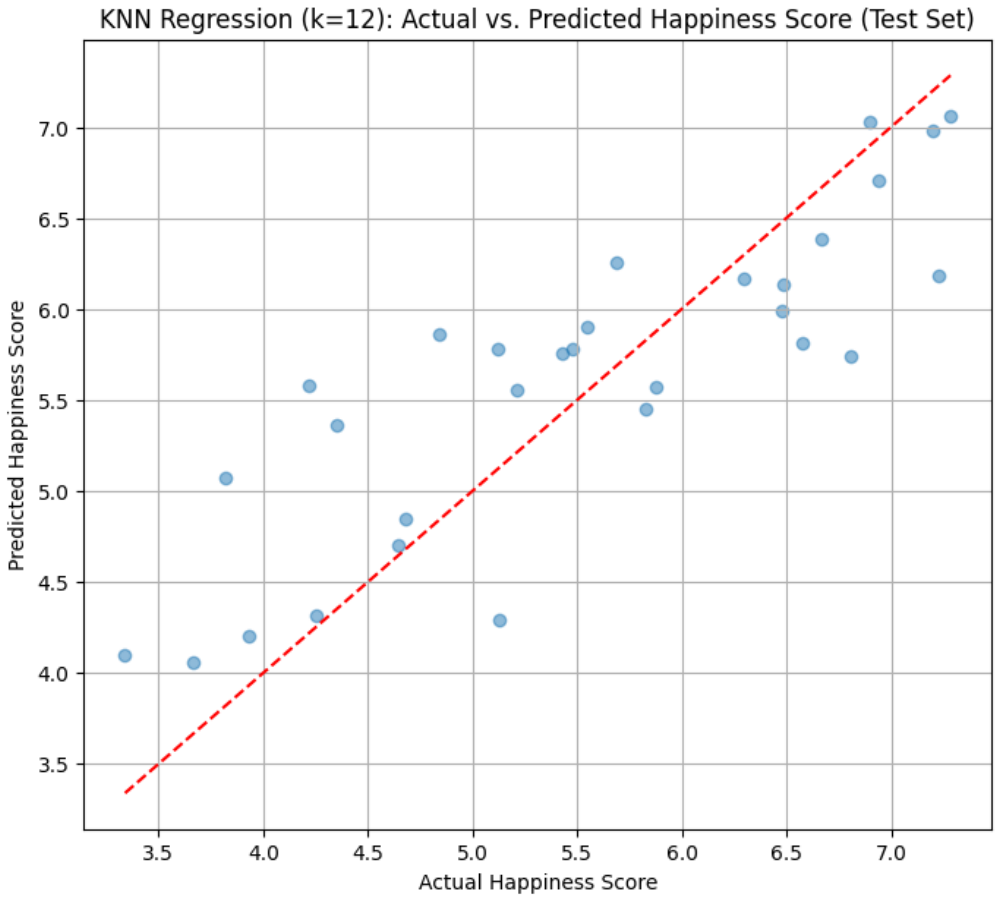
After confirming that KNN was the most suitable model, I wanted to visualize its predictions using the optimal k, so I applied cross-validation to identify the best neighbor count. Cross-validation was chosen because it provides a robust assessment of performance across multiple data splits, reducing overfitting and ensuring the selected k generalizes well.



Minimum Average Cross-Validated MSE: 0.9103

Best k: 12

Here is the improved-KNN with best k=12



Conclusion

In summary, this project has shown that while absolute life expectancy and key socio-economic factors—particularly freedom, GDP per capita, health, and generosity—strongly drive cross-national differences in happiness (with our best KNN model explaining ~70% of the variance at an RMSE of ~0.63), the gender gap in longevity plays a virtually negligible role. Hierarchical clustering further validated three coherent country groupings (high-income/high-happiness, large developed/resource-oriented, and emerging markets), and the integrated use of correlation testing, clustering, and optimized regression modeling provides a

clear, interpretable framework for understanding the complex determinants of national well-being.