Advanced Programming 2025

Explaining and Clustering Global Happiness

Final Project Report

Julia Jasinska
julia.jasinska@unil.ch
Student ID: 20348678

January 3, 2026

Abstract

In this project I analyze cross-country differences in happiness using data from the 2024 World Happiness Report. The dataset provides an average life evaluation ("ladder score") for 143 countries and six explanatory factors: log GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption. I first clean the data, standardize the six drivers and examine their correlation structure. This shows that GDP, health, social support, freedom and corruption form a strong "development and institutions" block, while generosity is largely independent of these variables. I then apply K-Means clustering to the standardized factors, selecting the number of clusters using silhouette scores. For the chosen K, I construct cluster profiles based on mean factor values and relate cluster membership to average happiness. Finally, I regress happiness on log GDP per capita to quantify the part of happiness explained by income alone and extend this regression with a high-corruption dummy and interaction terms between GDP, social support and corruption. The results reveal three clearly interpretable clusters that differ both in happiness levels and in their development profiles and highlight countries that are significantly happier or less happy than their income level would predict.

Keywords: data science, Python, machine learning, World Happiness Report, life satisfaction, unsupervised learning, K-means, silhouette score, standardization, correlation analysis, regression analysis, PCA vizualization

Contents

1 Introduction

Over the last decade, the measurement of well-being has moved beyond purely economic indicators. While gross domestic product per capita remains the standard yardstick for comparing countries, policymakers and researchers increasingly emphasize that income alone cannot fully capture how well people's lives are going. The World Happiness Report (WHR) responds to this concern by using survey data to construct a measure of "life evaluation" and by analyzing how economic, social and institutional factors jointly shape national happiness levels. [2]

In the 2024 edition, the WHR reports an average life evaluation ("ladder score") for 143 countries, alongside six explanatory variables: log GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption [2]. These six factors are not combined into a composite index; instead, they are used to explain why some countries score higher on the happiness ladder than others. Cross-sectional correlations suggest that GDP and other "development" variables are strongly associated with life evaluations, but they also leave open several questions. How similar are countries in terms of their full happiness profiles? Are there distinct groups of countries that share common socio-economic and institutional characteristics? And to what extent can national happiness be predicted by income alone versus non-economic factors?

This project addresses these questions using data-science tools on the WHR 2024 cross-country dataset. First, I characterize the structure of the six drivers by cleaning the data, standardizing the factors and studying their correlation matrix and principal components. Second, I apply the K-Means clustering algorithm to the standardized drivers to identify groups of countries with similar happiness profiles, selecting the number of clusters using silhouette scores and relating the resulting profiles to average ladder scores. Third, I quantify the role of GDP relative to the other factors using linear regressions: I regress happiness on log GDP per capita to obtain "happiness beyond GDP" residuals and then estimate an extended model with the six drivers, a high-corruption dummy and interaction terms between GDP, social support and corruption. In line with the "beyond GDP" agenda, GDP plays a central role in the analysis but is explicitly examined together with other drivers of happiness, rather than treated as a sufficient proxy for national well-being.

2 Literature Review / Related Work

Gross domestic product (GDP) per capita is the standard macroeconomic indicator used to compare countries, and it is often treated as a proxy for overall welfare. However, a large literature argues that GDP is an incomplete and sometimes misleading measure of well-being. The Stiglitz-Sen-Fitoussi Commission, set up by the French government, famously concluded that GDP primarily captures market production and "does not account for the quality of life in a comprehensive way," calling for a shift toward multidimensional measures that include health, education, social connections, subjective well-being and environmental sustainability. [3] This "beyond GDP" agenda has since been taken up by international organizations, which develop dashboards of indicators rather than a single income measure to assess social progress and guide policy.

A related empirical debate concerns how strongly GDP is actually linked to people's self-reported happiness. Early work by Easterlin argued that, while richer individuals within a country are consistently happier than poorer individuals, increases in a country's average income over time do not necessarily translate into higher average happiness. [4] This "Easterlin paradox" challenges the idea that long-run economic growth automatically raises national well-being and has motivated a large literature re-examining the income–happiness relationship. More recent analyses using larger and more comparable datasets find a robust positive association between log GDP per capita and average life satisfaction both across countries and over time, with no

clear saturation point at high income levels. [5] At the same time, these studies emphasize that income explains only part of the variation in well-being and that non-economic factors remain important.

The World Happiness Report (WHR) can be seen as a practical implementation of the "beyond GDP" perspective. Using Gallup World Poll data, it models cross-country differences in average life evaluation using six drivers: log GDP per capita, healthy life expectancy, social support, freedom to make life choices, generosity and freedom from corruption. [6] These variables together account for more than three quarters of the global variation in national happiness scores. In other words, GDP is a powerful but not sufficient predictor: countries with similar income levels can have very different happiness outcomes depending on their social and institutional context.

My project is directly situated in this literature. I take the 2024 World Happiness dataset and treat the ladder score as a measure of national life evaluation, while the six WHR drivers serve as a multidimensional description of each country's economic, social and institutional conditions. First, I use log GDP alone to explain cross-country happiness and compute residual happiness, which captures how much better or worse a country does relative to its income level. Second, I use clustering (k-means) and principal component analysis to group countries by their full profile of drivers, rather than by GDP alone. Finally, I estimate an extended regression model with interactions between GDP, social support and corruption, to study how economic and institutional dimensions jointly shape well-being. In this way, the project follows the "beyond GDP" agenda: GDP plays a central role, but it is explicitly analyzed alongside other drivers of happiness instead of being treated as a sufficient proxy for national well-being.

3 Data source and pre-processing

3.1 Data source

The analysis is based on the World Happiness Report 2024 dataset. The raw file ("*world-happiness-2024.csv*") contains country-level data with an average life evaluation ("*ladder score*") and eleven additional variables for 143 countries.

A first exploratory script (`src/explore_data.py`) loads the raw CSV, prints the shape of the dataset and lists all columns names. This step is mainly to check that the file is read correctly and to confirm the exact variable names before further processing.

3.2 Variables & Cleaning

In „*src/prepare_ data.py*" I build a simple dataset focused on my research questions. From the raw file I keep only:

- Country name- country identifier
- Ladder score - average national life evaluation on a 0-10 scale
- Six explanatory variables reported as contributions "*Explained by:*":
  - "*Explained by: Log GDP per capita*": log-transformed GDP per capita, contribution of economic prosperity.
  - "*Explained by: Social Support*": measure of having someone to count on in times of troubles.
  - "*Explained by: Healthy life expectancy*": proxy for health quality, access to healthcare, and longevity.

– "*Explained by: Freedom to make life choices*": perceived freedom of choice and control over life.

– "*Explained by: Generosity*": tendency to donate to charity and to help others.

– "*Explained by: Perceptions of corruption*": perception that corruption is low in government and business.

The remaining variables (upperwhisker, lowerwhisker, Dystopia+residual) are used in the original report but they are not part of my explanatory feature set so they are dropped. I also drop all rows with missing values in any of the six explanatory factors. This removes 3 countries from the original 143, leaving a final sample of 140 countries with complete information.

In the rest of the analysis, Ladder score is treated as the outcome variable: it is not used to construct clusters (to avoid clustering directly on happiness) but is used later to compare average happiness across clusters and as the dependent variable in the regression models.

Example code snippet:

```python
import pandas as pd

# Path to the original 2024 dataset
DATA_FILE = "data/world-happiness-2024.csv"
OUT_FILE = "results/clean_happiness_data.csv"


def main():
    print(">>> PREPARE_DATA.PY IS RUNNING <<<")

    # Load the full dataset with correct separators
    df = pd.read_csv(DATA_FILE, sep=";", decimal=",")

    # Select the variables needed for the project
    selected_columns = [
        "Country name",
        "Ladder score",  # happiness score (NOT used in clustering)
        "Explained by: Log GDP per capita",
        "Explained by: Social support",
        "Explained by: Healthy life expectancy",
        "Explained by: Freedom to make life choices",
        "Explained by: Generosity",
        "Explained by: Perceptions of corruption",
    ]

    # Keep only these columns
    clean_df = df[selected_columns].copy()

    # Show the shape to verify it worked
    print("Clean dataset shape BEFORE dropping missing values:", clean_df.shape
    )

    # Drop any country with missing values in these columns
    rows_before = len(clean_df)
    clean_df = clean_df.dropna(subset=selected_columns).copy()
    rows_after = len(clean_df)
```

Listing 1: Example data cleaning and selection

3.3 Transformations

The World Happiness Report already provides a variable called "Explained by: Log GDP per capita", which is a logarithmic transformation of GDP per capita. I use this component di-

rectly and rename it to "*log_ GDP*" for convenience. Working with log GDP rather than raw GDP levels is standard in happiness research: income is highly skewed across countries and the marginal effect of income on well-being is generally assumed to be diminishing. A log transformation compresses very high incomes and produces a more symmetric distribution, which is more appropriate for linear regression and for distance-based methods such K-Means.

3.4 Descriptive statistics

The scripts "src/standardize_data.py" and "explore_factors.py" focus on the six explanatory variables. Starting from the cleaned dataset, I first rename the original labels to shorter, more readable names:

- "*Explained by: Log GDP per capita*" → "*log_ GDP*"
- "*Explained by: Social Support*" → "*Social_ Support*"
- "*Explained by: Healthy life expectancy*" → "*Life_ expectancy*"
- "*Explained by: Freedom to make life choices*" → "*Freedom*"
- "*Explained by: Generosity*" → "*Generosity*"
- "*Perceptions of corruption*" → "*Corruption*"

Using the standardized version of these six drivers (constructed in Section 4.1), I compute descriptive statistics (mean, standard deviation, minimum, maximum) for each factor and save them to *results/factor_ summary.csv*. I also compute the correlation matrix of the six drivers (*results/factor_ correlations.csv*) and visualize it as a heatmap (Appendix, *Figure 1*). The heatmap shows a clear block of strong positive correlations between *log_ GDP*, *Social_ Support*, *Life_ expectancy*, *Freedom*, and *Corruption*, while *Generosity* is only weakly correlated with the other variables. This already suggests one dominant "development and institutions" dimension plus a more independent generosity dimension, which later helps interpret both the PCA and the K-Means clusters.

4 Methodology

All analyses are implemented in Python (JupyterLab 3.3.4) using pandas, scikit-learn and matplotlib. The pipeline consists of: (i) feature engineering and standardization, (ii) unsupervised clustering with K-Means, (iii) dimensionality reduction with Principal Component Analysis (PCA), (iv) linear regression models, including interactions, and (v) simple testing and robustness checks.

4.1 Feature engineering

Most of the modelling in this project (K-Means clustering and principal component analysis) relies on Euclidean distances or on variance-based decompositions. To avoid giving disproportionate weight to variables with large scales or variances, I standardize the six explanatory variables using StandardScaler from scikit-learn with default hyperparameters (with_mean=True, with_std=True). For each factor x_j, I construct a z-score

$$x_{j,std} = \frac{x_j - \bar{x}_j}{sd(x_j)}, \tag{1}$$

so that each standardized feature has mean 0 and standard deviation 1. The standardized dataset is saved to *results/happiness_ standardized.csv*, and it is used as the input for K-Means and PCA.

To verify that standardization worked as intended, I include a small test ("*tests/test_ standardization.py*") which checks that each standardized factor has mean close to zero and standard deviation close to one (up to a small numerical tolerance).

To capture institutional non-linearity, I define a *HighCorruption* dummy equal to 1 for countries in the bottom 30% of the corruption distribution (i.e. perceived as most corrupt) and 0 otherwise. For the extended regression model, I also construct two interaction terms, *logGDP×SocialSupport* and *logGDP×HighCorruption*, and standardize them before estimation so that regression coefficients remain comparable in magnitude. The correlation heatmap of the six drivers is documented separately in Section 3.4 (Appendix, *Figure 1*).

4.2 Clustering with K-Means

The main unsupervised learning method in this project is K-Means clustering applied to the six standardized drivers described above for 140 countries with complete data. The goal is to group countries into homogeneous clusters based on their profiles, without using the happiness score itself to form clusters. In practice, I use scikit-learn's "K-Means" with Euclidean distance, the default maximum number of iterations, ten random initialization ("n_Init = 10") and a fixed random seed ("random_state = 42") to obtain a stable and reproducible clustering. The main modelling choice is the number of clusters. To choose the number of clusters, I compute the average silhouette score for K = 3,4,5,6. For each country i, the silhouette score is defined as:

a(i): average distance from i to all other points in its own cluster,

b(i): smallest average distance from i to points in any other cluster,

and

$$x_{j,std} = \frac{x_j - \bar{x}_j}{sd(x_j)}, \tag{2}$$

The index summarizes, for each country, how close it is to other countries in its cluster compared with countries in the nearest alternative cluster. The final score is the average across all observations. Values close to 1 indicate well-separated, compact clusters, values around 0 indicate overlapping clusters, and negative values suggest that some points would be closer to another cluster. In my application, the silhouette score is highest for K = 3, with only modest gains or even slight deteriorations for larger K. I, therefore, retain three clusters as the best compromise between model fit and interpretability.

4.3 PCA and regression models

To obtain an interpretable two-dimensional representation of the data, I complement the K-Means analysis with Principal Component Analysis (PCA). I fit PCA to the standardized feature matrix and retain the first two components to obtain a two-dimensional representation. I then project each country onto the resulting PC1–PC2 space and plot the scores, coloring points by their K-Means cluster assignment. To interpret the axes, I rely on the component loadings (which indicate which drivers contribute most to each principal component) and on how clusters distribute in this reduced space.

To quantify how strongly happiness is related to income and other drivers, and to interpret why clusters differ, I estimate two types of linear regression models using scikit-learn's LinearRegression (with default hyperparameters, in particular fit_intercept=True). In the first model, the dependent variable is the happiness ladder score, and the only predictor is log GDP per capita.

$$Ladder\ score_i = \alpha + \beta \log GDP_i + \varepsilon_i. \tag{3}$$

I fit this model on all 140 countries using "*src/gdp_ residuals.py*". This model allows me to measure how much of the cross-country variation in happiness can be explained by income alone. The residuals from this regression are interpreted as "happiness beyond GDP" and are

examined across clusters to see whether some groups of countries are systematically above or below the income - happiness line (Appendix, *Figure 3*).

In the second model, I estimate an extended multiple regression to study how economic, social and institutional factors interact:

$$\text{Ladder score}_i = \alpha + \sum_j \beta_j X_{ij} + \gamma_1 \text{HighCorruption}_i + \gamma_2 \left(\log\_GDP_i \times \text{Social\_Support}_i\right)$$

$$+ \gamma_3 \left(\log\_GDP_i \times \text{HighCorruption}_i\right) + u_i, (4)$$

where X_ij includes the six standardized drivers (log GDP, social support, life expectancy, freedom, generosity, corruption).

This specification is included in "*src/reginteractions.py*". Using standardized predictors allows a direct comparison of coefficient magnitudes and helps assess whether the happiness returns to income are higher in countries with strong social support and low corruption.

4.4 Testing and robustness

To check that the pipeline behaves as intended and to address robustness concerns, I implement two types of checks. Firstly, using simple tests (in the "*tests/*" folder), I verify that:

- the standardized drivers have mean approximately zero and standard deviation approximately one (with a small numerical tolerance),
- there are no missing values in the standardized features used for K-Means,
- among K 3, 4, 5, 6, the best silhouette score is indeed obtained at K = 3,
- the first two principal components explain at least 60% of the total variance in the six drivers, and
- the final K-Means solution for K = 3 produces exactly three distinct clusters.

These tests are not exhaustive, but they provide basic confidence that the main preprocessing and clustering steps are implemented correctly.

Secondly, because GDP plays a dominant role among the drivers, I re-estimate the K-Means model without the log GDP factor in "*src/run_kmeans_no_gdp.py*", keeping the same number of clusters (K=3). I then compare the baseline cluster labels (with GDP) to the alternative labels (without GDP) using a simple confusion matrix saved in "*results/cluster_confusion_no_gdp.csv*" and summarized in "*src/compare_clusters.py*". This robustness exercise checks whether the high- and low-happiness groups are driven almost entirely by income, or whether the clustering structure is preserved when GDP is removed and clusters are formed mainly based on social support, health, freedom, generosity and corruption.

Together, these methods form the core of the data-science and machine-learning component of the project: unsupervised learning (K-Means and PCA) to discover structure in the drivers of happiness, and supervised learning (linear regression with interactions) to quantify how different factors and their combinations are associated with national well-being.

5 Results

5.1 Descriptive patterns and correlations

I start by inspecting how the six explanatory factors co-move across countries. *Figure 1* shows a correlation heatmap of the standardized variables (log GDP, social support, healthy life expectancy, freedom, generosity and perceptions of corruption). All pairs related to broad "development and institutions" are strongly and positively correlated: log GDP correlates around 0.7–0.8

with social support and life expectancy, and around 0.4–0.5 with freedom and the corruption measure. In contrast, generosity is only weakly related to the other factors, with correlations close to zero. This pattern suggests that most factors move together along a general development axis, while generosity behaves more independently. This motivates both the clustering step and the later interaction analysis.

5.2 K-Means clusters and factor profiles

Using the six standardized factors as input, I run K-Means with K=3. This partitions the 140 countries into three groups of unequal size: cluster 0 contains 73 countries, cluster 1 contains 46 countries, and cluster 2 contains 21 countries. Average happiness levels differ markedly between these groups. The disadvantaged cluster (cluster 1) has the lowest mean ladder score (about 4.25), the intermediate cluster (cluster 0) has a higher average of 5.93, and the high-development cluster (cluster 2) reaches 6.94. Moving from the low- to the high-development cluster is thus associated with an increase of almost three points on the 0–10 happiness scale.

Figure 6 shows the mean z-scores of the six factors in each cluster. Cluster 1 (orange) scores well below zero on log GDP, social support, life expectancy and freedom, and has slightly lower perceptions of corruption than the world average, which fits a group of poorer and less stable countries. Cluster 2 (green) has clearly positive scores on all six variables, especially perceptions of corruption, which is roughly two standard deviations above the global mean. This cluster corresponds to highly developed, well-governed countries with strong social support, long lives and high freedom. Cluster 0 (blue) lies between these two extremes, with moderately above average income and social indicators but weaker institutions and generosity. The pattern confirms that the clustering essentially recovers a development/institutions axis, while generosity varies more weakly across clusters.

To assess how much the clustering depends on GDP, I re-estimate K-Means on the five non-GDP factors (saved in *results/cluster_assignments_no_gdp.csv*) and compare the new labels with the baseline (saved in *results/cluster_confusion_no_gdp.csv*). Most countries remain in the same group; only eight change cluster. The reassignments are informative: some countries, such as France, drop from the high-development cluster to the intermediate one when GDP is removed, indicating that their position among the richest group is driven partly by income rather than by uniformly strong social and institutional scores. Others, such as Venezuela and Laos, move up when clustering is based on non-income drivers alone, suggesting that they perform better on social or institutional dimensions than their income level would suggest. This robustness check confirms that GDP is important but not the only factor shaping the cluster structure.

5.3 PCA visualization of the clusters

PCA shows that the first two components capture most of the information in the six drivers: PC1 explains 51.3% of total variance and PC2 explains 19.2%, for a combined 70.5%. In the PC1–PC2 plot (*Figure 2*), the clusters separate mainly along PC1, which aligns with a broad "development and institutional quality" gradient: cluster 2 lies predominantly on the right (higher GDP, social support, life expectancy, freedom, and lower corruption), cluster 1 concentrates on the left (lower values on the same bundle), and cluster 0 falls in between. PC2 adds secondary separation, capturing more country-specific differences that are less tightly connected to income and institutions, including variation related to generosity. Overall, the figure suggests the clusters are ordered along a clear continuum from low to high development rather than forming unrelated groups.

5.4 Happiness vs GDP

A central question in this project is how much of the cross-country variation in happiness can be explained by income alone. To address this, I regress the ladder score on log GDP per capita. The estimated slope is about 2.1, meaning that in this sample a one-unit increase in the log-GDP index is associated with roughly a two-point increase in average happiness on the 0–10 scale. The regression explains around 59% of the variance in happiness ($R^2 \approx 0.59$), so income is an important but clearly not sufficient determinant of national well-being; the fitted line in Figure 3 summarizes this baseline relationship.

However, *Figure 3* also shows substantial dispersion around the line. Some countries are roughly three ladder points happier than predicted by GDP alone, while others are almost three points less happy. When I color points with a cluster, a clear pattern emerges. Countries in cluster 1 (the disadvantaged cluster) are typically located below the regression line: given their income level, they are less happy than the model would predict. Examples include Sri Lanka, Botswana and Eswatini, which combine modest GDP with relatively low happiness. In contrast, most cluster-2 countries (high-development democracies such as Finland, Denmark and the other Nordic countries) lie above the line and are happier than predicted by GDP alone. Countries in cluster 0 tend to lie closer to the fitted line, with a mix of small positive and negative deviations. These results suggest that social and institutional factors related to support, health, freedom and corruption systematically influence how effectively income is translated into happiness.

To quantify this pattern, I compute residuals from the regression (actual minus predicted happiness) and compare them across clusters. *Figure 8* plots residual happiness by cluster. Cluster 1 clearly has negative residuals on average, meaning these countries are less happy than their GDP would predict, while cluster 2 has mostly positive residuals and cluster 0 is centered around zero. This confirms that the clusters do not only differ in income levels, but also in how effectively income and institutions are converted into subjective well-being.

5.5 Regression with interactions (*Figure 7*)

To better understand why clusters differ, I estimate a multiple linear regression of happiness (Ladder score) on the six WHR factors, enriched with a high-corruption dummy and interaction terms: log GDP x social support and log GDP x high corruption. All predictors are standardized before estimation, so the coefficients can be compared in magnitude.

This model explains 83% of the variance in happiness ($R^2 = 0.83$, RMSE ≈ 0.49). The largest coefficient is on the interaction between log GDP per capita and social support ($\beta \approx 0.75$), indicating that economic development translates into much higher happiness when it is accompanied by strong social ties. Freedom to make life choices ($\beta \approx 0.32$) and healthy life expectancy ($\beta \approx 0.25$) also have sizeable positive effects. The dummy for highly corrupt countries has a negative coefficient ($\beta \approx -0.18$), suggesting that corruption depresses well-being even after controlling for income and other factors. The main effect of log GDP is slightly negative, but it must be interpreted jointly with the interaction term: for countries with above-average social support, the combined effect of higher GDP is strongly positive. Overall, the regression with interactions supports the descriptive cluster results: beyond GDP, high levels of social support, freedom, health and low corruption contribute jointly to higher happiness, and the benefit of social support is especially strong in richer, less corrupt settings.

5.6 Additional interaction visualizations

Finally, I complement the main results with additional visualizations that illustrate how pairs of drivers interact with GDP and with each other.

First, *Figure 4* plots log GDP per capita against healthy life expectancy, with points colored by K-Means cluster. The cloud is close to a straight line: richer countries almost always enjoy

longer healthy lives, and the three clusters are clearly ordered along this gradient. Cluster-1 countries occupy the bottom-left region with low GDP and short life expectancy; cluster-0 countries are in the middle; and cluster-2 countries are concentrated in the top-right corner. This plot shows that the "development" axis is not driven by GDP alone but by a tight link between economic resources and public-health outcomes.

Second, *Figure 5* shows log GDP per capita against generosity. Here the picture is very different: the scatter is much more diffuse, with only a very weak negative correlation and a lot of overlap between clusters. Poor countries can be relatively generous, while some rich countries are less so, and vice versa. This confirms the impression from the heatmap that generosity is only loosely related to the development bundle and captures a more independent cultural dimension. Together with the regression results, these additional plots highlight that some drivers of happiness (income, health, freedom, corruption) move in a tightly coordinated way, while others (generosity) provide extra variation that is not explained by GDP alone.

6 Discussion & Limitations

The analysis of the 2024 World Happiness data reveals a relatively clear structure behind cross-country differences in life evaluation. Clustering, PCA and regression all point to one dominant "development and institutions" dimension - combining income, social support, healthy life expectancy, perceived freedom and low corruption - and a secondary, more independent generosity dimension. Along this structure, K-Means identifies three groups of countries: a high-development cluster with favorable scores on almost all drivers and the highest happiness; an intermediate group with moderate conditions and intermediate happiness; and a disadvantaged cluster with low income, poorer health, weaker freedoms, higher corruption and the lowest ladder scores.

The first key finding is that GDP per capita is a strong but incomplete predictor of national happiness. The simple regression of the ladder score on log GDP explains a large share of the cross-country variation and the scatter plot shows a clear positive association between income and life evaluation. At the same time, residuals from this regression reveal substantial "happiness beyond GDP": countries in the high-development cluster tend to lie above the regression line, while those in the disadvantaged cluster lie below it. This suggests that social and institutional factors systematically affect how effectively income is converted into subjective well-being.

The extended regression model with standardized predictors and interaction terms clarifies these mechanisms. Once all six drivers, the high-corruption dummy and the interaction terms are included, the explanatory power of the model increases markedly. The largest coefficient is on the interaction between log GDP and social support, indicating that high income and strong social networks reinforce each other: economic development is associated with especially high happiness when people also report having someone to rely on. The high-corruption dummy enters with a negative coefficient, and the interaction between log GDP and high corruption suggests that the happiness returns to income are weaker in more corrupt environments. These results complement the cluster profiles and PCA plot: the happiest countries are those that combine high income with strong support, good health, high perceived freedom and relatively low corruption, while countries with similar income but weaker institutions underperform.

Despite these insights, the analysis has several limitations. First, the data are cross-sectional: each country appears only once, in 2024. The models therefore capture associations, not causal effects. It is not possible to conclude from this project that increasing GDP, improving social support or reducing corruption would mechanically cause a given change in happiness. The results are descriptive and should be interpreted as such. A panel dataset with multiple years per country would be better suited to studying dynamic or causal relationships between changes in drivers and changes in well-being. Second, the variables are partly based on survey responses and on the World Happiness decomposition, so measurement error and cultural differences in how

people answer questions may affect the estimates. Third, the modelling choices, like K-Means with Euclidean distance and linear regressions estimated on the full sample, are deliberately simple. More flexible clustering algorithms or non-linear models such as Gaussian mixture models or hierarchical clustering could capture different patterns. Finally, strong correlations between income, support, health, freedom and low corruption make it difficult to cleanly separate their individual effects, and the analysis is restricted to the six drivers available in the 2024 report, leaving out other potential determinants such as inequality, unemployment or environmental quality.

These limitations suggest several directions for future work. Using panel data from multiple years would allow an analysis of how countries move between clusters over time, and whether improvements in institutions or social support precedes increases in happiness. Extending the set of predictors to include additional economic, social and environmental indicators could refine the classification of country happiness regimes. Methodologically, it would be interesting to compare K-Means with more flexible clustering algorithms and to use regularized or non-linear models (such as random forests) to assess the robustness of the regression results. Within the scope of this project, however, the combination of clustering, PCA and linear regression already provides a coherent and interpretable picture of how income, social support, health, freedom, generosity and corruption jointly relate to cross-country differences in reported well-being.

7 Conclusion

This project uses the 2024 World Happiness Report to study how economic, social and institutional factors relate to cross-country differences in life evaluation. Starting from the WHR framework, I focus on log GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption, and apply a simple but coherent data-science workflow: cleaning and standardizing the data, clustering countries with K-Means, reducing dimensionality with PCA, and estimating linear regressions, including interaction terms.

The results show a clear pattern. Log GDP per capita is a strong predictor of national happiness, but it does not fully determine it: the GDP-only regression leaves substantial residual variation, and K-Means identifies three clusters with distinct profiles of development and institutions and markedly different average ladder scores. PCA reveals that a single "development and institutions" dimension accounts for most of the variation in the six drivers, while generosity behaves more independently. The extended regression with standardized predictors confirms that happiness is highest in countries that combine high income with strong social support, good health, high perceived freedom and relatively low corruption, and lower where institutional quality is weak. Within the scope of this course, these findings meet the main objective of the project: to use standard data-science and machine-learning methods to provide an interpretable, quantitative picture of the determinants of national happiness in 2024.

References

1. Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (Eds.). (2024). World Happiness Report 2024. University of Oxford: Wellbeing Research Centre. Available at: https://www.worldhappiness.report/ed/2024/

2. Helliwell, J., Layard, R., Sachs, J. et al. (2025). Caring and Sharing: Global Analysis of Happiness and Kindness, in World Happiness Report 2025. Sustainable DevelopmentSolutions Network. Available at: https://www.worldhappiness.report/ed/2025/caring-and-sharing-global-analysis-of-happiness-and-kindness/

3. Stiglitz, J. E., Sen, A., & Fitoussi, J.-P. (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress. European Commission. Available at: https://ec.europa.eu/eurostat/documents/8131721/8131772/Stiglitz-Sen-Fitoussi-Commission pdf

4. Easterlin, R. A. (1974). "Does Economic Growth Improve the Human Lot? Some Empirical Evidence." In P. A. David & M. W. Reder (eds.), Nations and Households in Economic Growth. Academic Press. Available at: https://www.brookings.edu/wp-content/uploads/2008/03/2008a_bpea_stevenson.pdf

5. Stevenson, B., & Wolfers, J. (2008). "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox." Brookings Papers on Economic Activity, Spring 2008. Available at: https://ses.sp.bvs.br/local/File/High%20income%20improves%20evaluation%20of%20life%20but%20not%20emotional%20well-being.pdf

6. Kahneman, D., & Deaton, A. (2010). "High Income Improves Evaluation of Life But Not Emotional Well-Being." Proceedings of the National Academy of Sciences, 107(38), 16489–16493. Available at: https://files.worldhappiness.report/WHR24.pdf

A Additional Figures

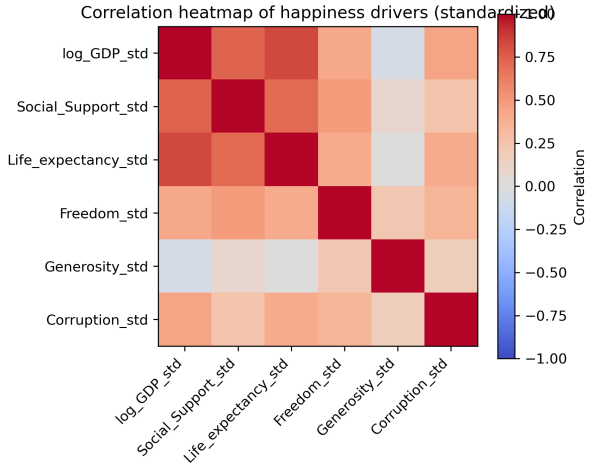

Figure 1: Correlation heatmap of the six standardized happiness drivers (log GDP, social support, healthy life expectancy, freedom, generosity and corruption)

Figure 2: PCA of happiness drivers with K-Means clusters (PC1 vs PC2, coloured by cluster)

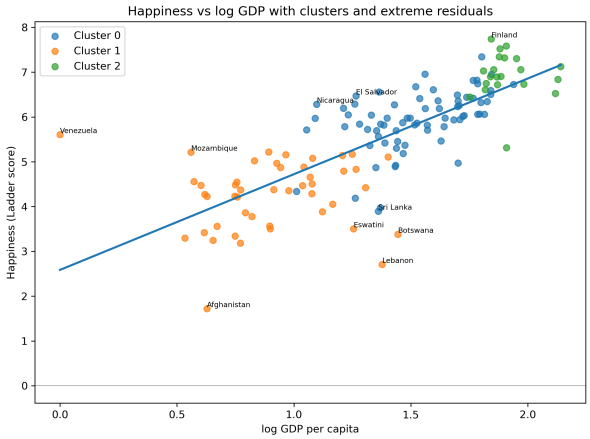

Figure 3: Happiness vs log GDP per capita with K-Means clusters and regression line (countries with extreme residuals labelled)

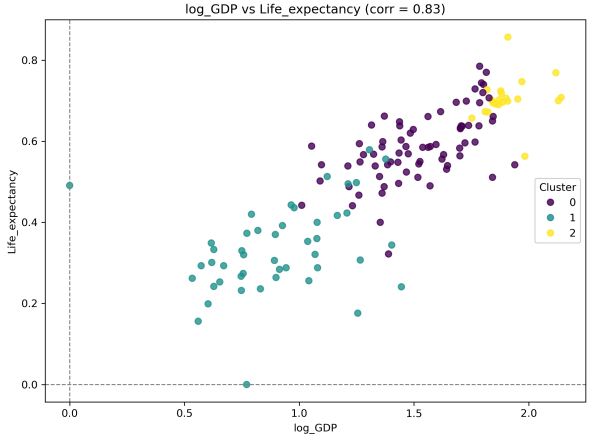

Figure 4: log GDP per capita vs healthy life expectancy by cluster

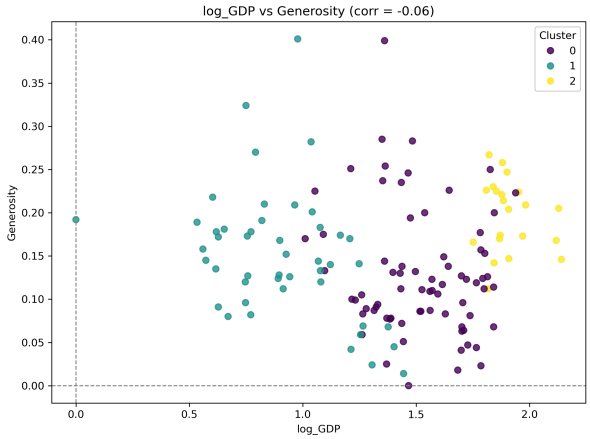

Figure 5: log GDP per capita vs generosity by cluster

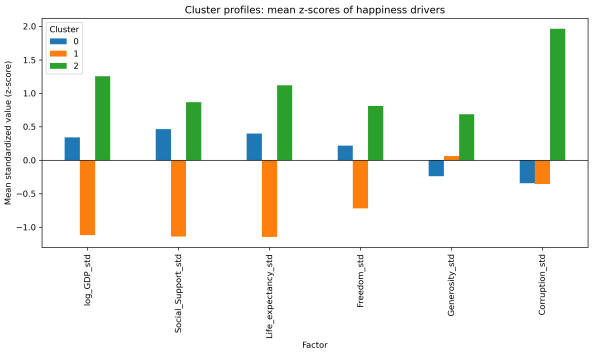

Figure 6: Cluster profiles: mean standardized factor values (z-scores) for each happiness driver

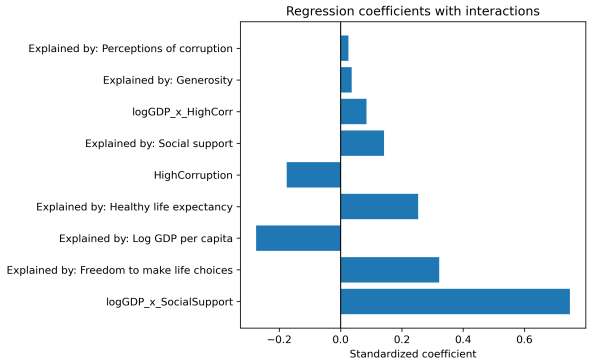

Figure 7: Standardized regression coefficients from the model with interactions and HighCorruption dummy

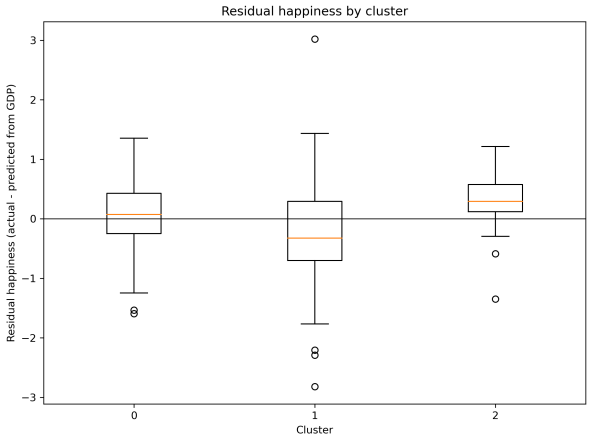

Figure 8: Residual happiness (actual – predicted from GDP-only model) by cluster (boxplot)

B Code Repository

GitHub Repository: `https://github.com/jwjasin-ctrl/datascience-happiness-project.git`

- Repository structure

  The repository is organized as follows:

  - `data/`: raw input data files (World Happiness Report 2024 dataset).
  - `src/`: Python scripts for data cleaning, exploratory analysis, clustering, PCA, and regressions.
  - `tests/`: simple tests (e.g., checking standardization).
  - `results/`: generated outputs (tables and figures) produced by the pipeline.
  - `requirements.txt`: Pip dependencies (recommended)
  - `environment.yml`: Conda dependencies (optional)
  - `README.md`: project overview and repository structure
  - `run_all.py`: to execute the full code

- Installation instructions (see in *README.md*)

```
git clone https://github.com/jwjasin-ctrl/datascience-happiness-project.git
cd datascience-happiness-project

python3 -m venv .venv
source .venv/bin/activate

python -m pip install --upgrade pip
pip install -r requirements.txt

python run_all.py
```

Listing 2: Installation instructions

- How to reproduce results

  Run the full pipeline: python run_all.py

C AI tools used

I used OpenAI's ChatGPT as a helper tool to clarify programming concepts, assist with debugging Python code, and improve the writing of this report (by correcting English and polishing sentence formulation). All final code was written, run, and checked by myself, and I made all modelling and interpretation decisions.