Data Science in Practice

Analysis of Socio-Economic Factors on World Happiness

Authors:

Aayush Randeep
Deep Jangle
Goutham Deepak
Sarthak Gaur
Shrutisagar Jena

Course Instructor: Dr. Samiran Das

1 Introduction

The World Happiness Report collects annual country-level data on various socio-economic and psychological factors contributing to happiness. The objective of this project was to merge the datasets from 2015 to 2019 into a single panel DataFrame suitable for analysis.

Because each year's dataset had slight differences in column names and formatting, the merging process required systematic preprocessing. Using Python and the `pandas` library, we ensured consistency across all datasets before combining them.

2 Step 1: Loading the Individual Datasets

We began by loading each CSV file into a separate pandas DataFrame using the `read_csv()` function. Example:

```
df_2015 = pd.read_csv('2015.csv')
df_2016 = pd.read_csv('2016.csv')
df_2017 = pd.read_csv('2017.csv')
df_2018 = pd.read_csv('2018.csv')
df_2019 = pd.read_csv('2019.csv')
```

This resulted in five DataFrames, each containing the happiness metrics for a specific year.

3 Step 2: Standardizing Column Names

The raw datasets were inconsistent in their column naming conventions. For instance:

- `Happiness Score` appeared as `Score` in some years.
- `Economy (GDP per Capita)` was listed as `GDP per capita` in later years.

To establish uniformity, we constructed a dictionary mapping each variant to a standardized name (e.g., "Happiness Score", "GDP per Capita"). Each DataFrame was then updated using:

```
df.rename(columns=column_mapping)
```

This ensured consistent naming across all five datasets.

4 Step 3: Selecting Common Columns

We identified the set of variables that appeared in all five datasets. These were:

- Country
- Happiness Rank
- Happiness Score
- GDP per Capita
- Social Support
- Life Expectancy
- Freedom
- Perceptions of Corruption
- Generosity

Each DataFrame was filtered to retain only these shared columns, ensuring structural compatibility for merging.

5 Step 4: Adding a Year Column

To preserve the panel nature of the dataset and identify each observation's year, we added a new column:

```
df_2015["Year"] = 2015
df_2016["Year"] = 2016
...
df_2019["Year"] = 2019
```

This enabled chronological analysis and comparisons across years.

6 Step 5: Concatenating the DataFrames

After standardization, the five DataFrames were vertically concatenated using:

```
merged_df = pd.concat(
    [df_2015, df_2016, df_2017, df_2018, df_2019],
    ignore_index=True
)
```

The resulting merged dataset contained approximately 790 rows, corresponding roughly to 158 countries per year across five years.

7 Step 6: Handling Missing Values and Data Types

To ensure the data was analysis-ready, we performed several cleaning operations:

- Replaced placeholders such as `"N/A"` with `NaN`.
- Imputed missing numerical values using the median.
- Converted numerical columns to `float` type.
- Removed duplicate entries based on the combination of `Country` and `Year`.

These steps improved consistency and prevented bias in downstream analysis.

8 Step 7: Verifying the Merge

We conducted several checks to validate the merged dataset:

- `df.info()` to verify data types.
- `df.head()` to inspect sample entries.
- `df.shape` to confirm total size.
- `df.isnull().sum()` to evaluate missing values.
- `df["Country"].nunique()` to ensure correct country coverage.

After verification, the dataset was saved as:

```
merged_df.to_csv('merged_happiness_data.csv', index=False)
```

9 Conclusion

Through systematic preprocessing and careful standardization, we successfully merged the World Happiness Report datasets from 2015 to 2019 into a unified panel dataset. The final DataFrame is fully cleaned, consistently structured, and suitable for exploratory analysis and modeling.

This merged dataset enables richer insights by allowing comparisons across countries and across time, laying the foundation for further statistical and machine learning analyses.

10 Outlier Removal

Outliers are data points which significantly differ from the general trend of the dataset. In order to understand better we need to remove them. They can affect the properties of the datasets like mean,variance and correlation values significantly.

When we try to do regression, the model can fit the dataset wrongly due to the presence of outliers, which in turn give us bad predictions.

The popular methods to remove them are as follows:

- **Z-Score** : We calculate the mean(μ) and the standard deviation(σ). Then we define z score as

  $$z = \frac{x - \mu}{\sigma}$$

  any points that have the

  $$|z| > 3$$

  are termed as outliers. This is especially helpful when the data follows normal distribution.

- **IQR Range** : Here we remove outliers based on the IQR(Inter Quartile Range). We have

  $$IQR = Q3 - Q1$$

  and

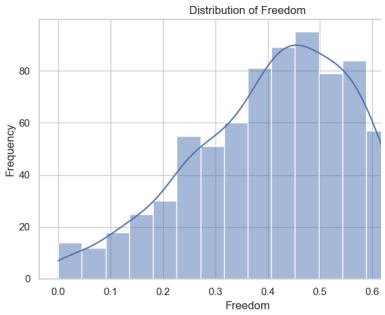
  $$x < Q1 - 1.5\ IQR \text{ and } x > Q3 + 1.5\ IQR$$

  are deemed as outliers.

- **Other Methods** : There are other methods like Isolation Forest , LOF – Local Outlier Factor method which are also used in detecting outliers.

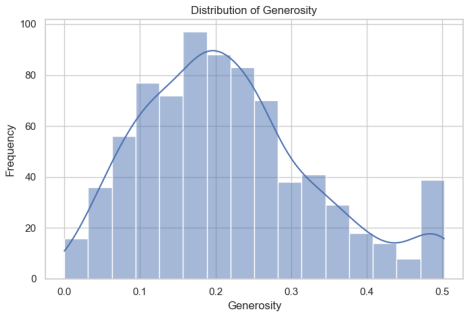
  In our analysis we will use the IQR Method for outlier detection and removal.

11 Distribution of the data after cleaning



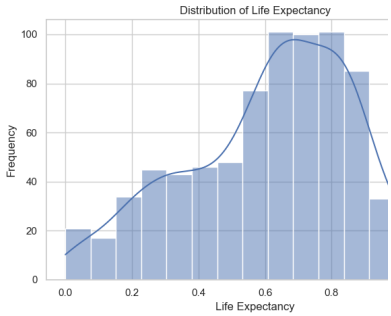
(a) Freedom

(b) GDP per Capita

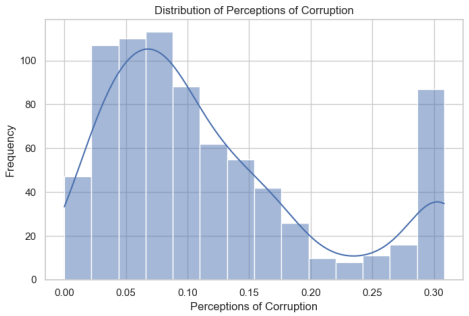


(c) Generosity

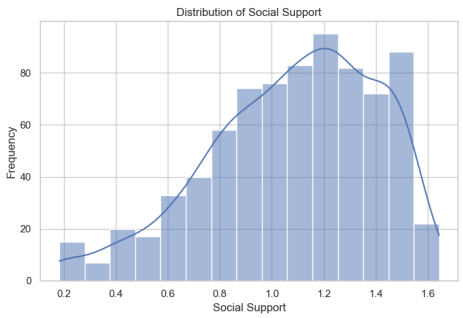
(d) Happiness Score



(e) Life Expectancy



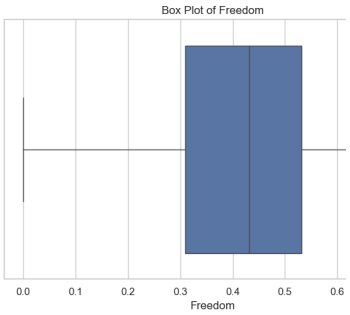
(f) Perceptions of Corruption



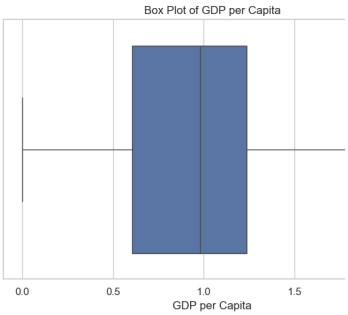
(g) Social Support

Figure 1: Distributions of cleaned variables (after capping).

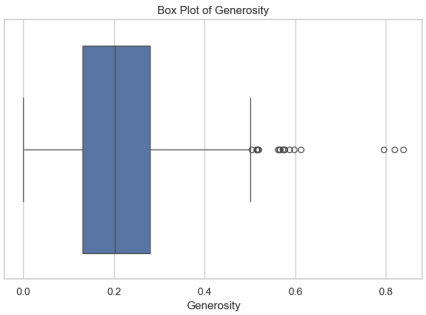
12 Box Plot after Cleaning the Data



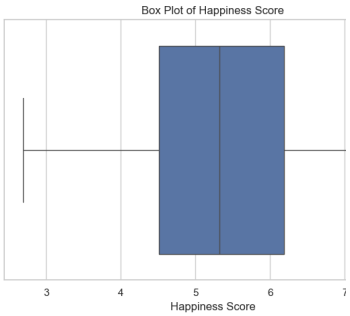
(a) Freedom



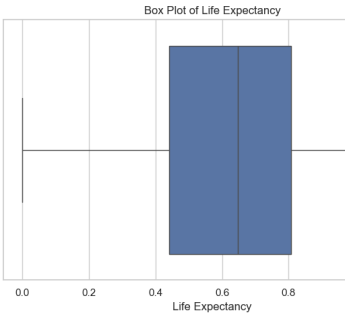
(b) GDP per Capita



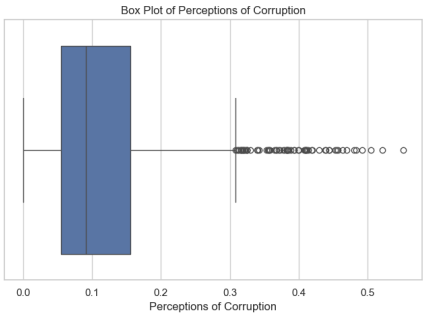
(c) Generosity



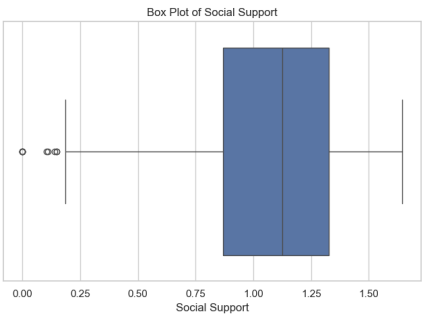
(d) Happiness Score



(e) Life Expectancy



(f) Perceptions of Corruption



(g) Social Support

Figure 2: Box plots of cleaned variables (capped at IQR bounds).

12.1 Correlation Analysis

To explore how different socio-economic variables are associated with national happiness, we first computed a correlation matrix and visualised it using a heatmap. This exploratory analysis helps identify which features have the strongest linear relationships with the Happiness Score and therefore deserve more attention in the modelling phase.

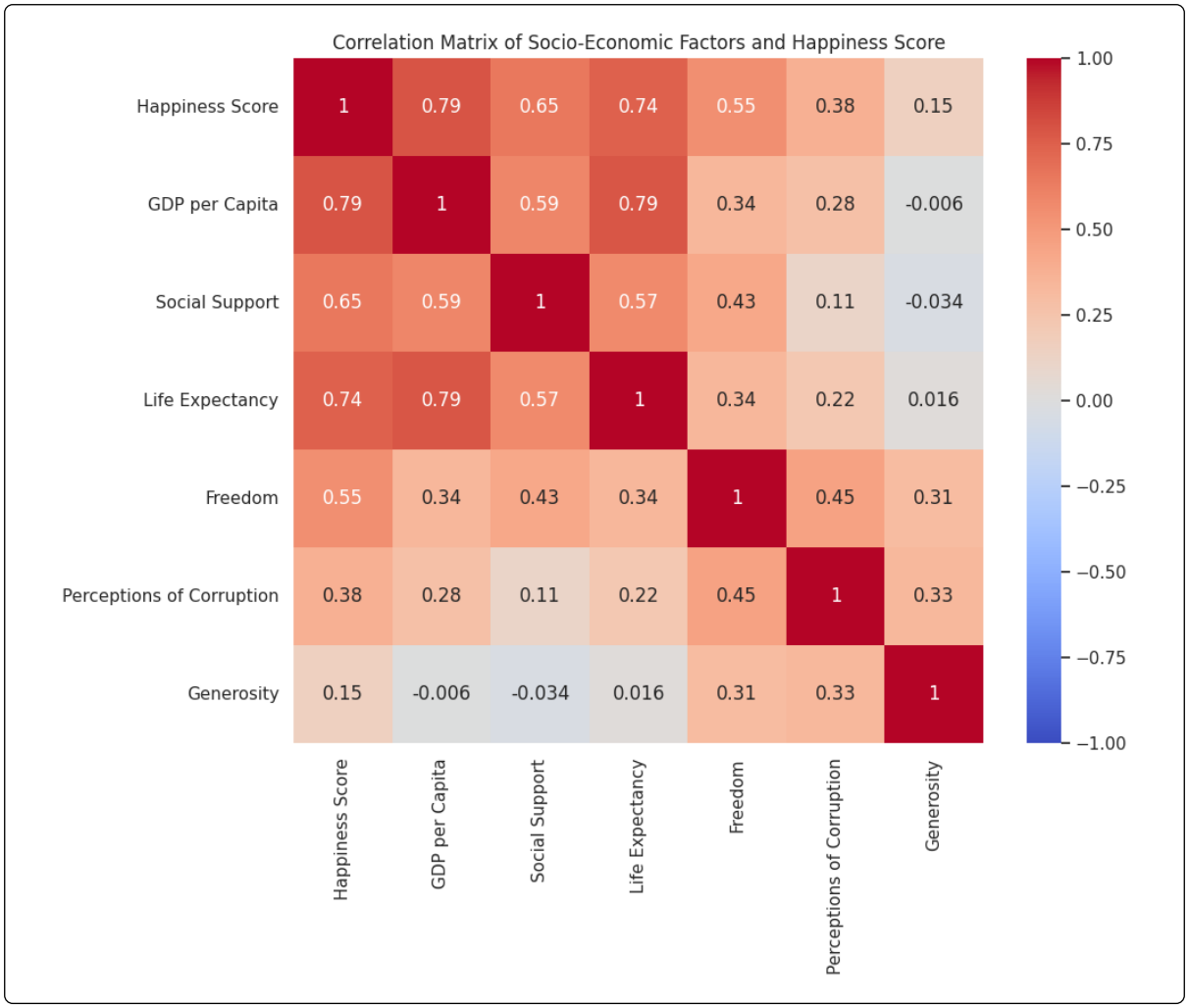


Figure 3: Correlation heatmap of socio-economic variables and Happiness Score.

From the heatmap, several important patterns emerge. The three factors that exhibit the strongest positive correlation with the Happiness Score are:

- **GDP per Capita** ($r = 0.79$)
- **Life Expectancy** ($r = 0.74$)
- **Social Support** ($r = 0.65$)

These results suggest that economic prosperity, overall health, and the strength of social networks play a central role in determining the happiness levels of a country. Countries with higher income levels, longer life expectancy, and stronger community support systems tend to report higher happiness scores — a trend that is consistent with findings from international socio-economic studies.

We also observe moderate correlations with:

- **Freedom to make life choices** ($r = 0.55$)

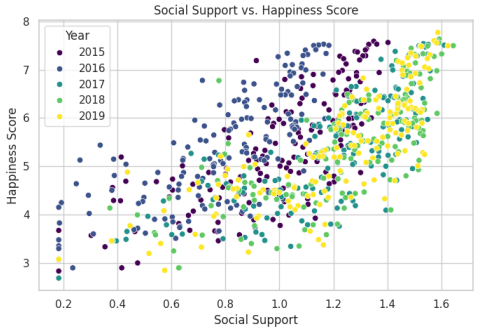
- **Perceptions of Corruption** ($r = 0.38$)

Although these correlations are weaker than the top three predictors, they still indicate meaningful relationships. Greater freedom generally enhances well-being, while higher corruption tends to reduce trust and satisfaction, which can negatively affect happiness.

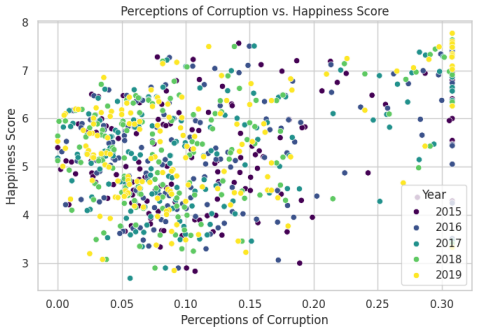
Finally, the variable **Generosity** shows a relatively weak correlation with happiness ($r = 0.15$). This implies that while generosity may contribute to personal well-being, it does not strongly influence national-level happiness scores.

Overall, the correlation heatmap provided essential insights into the underlying structure of the dataset. It helped us identify which variables are most relevant for predicting happiness and guided the selection of features used in our regression modelling.

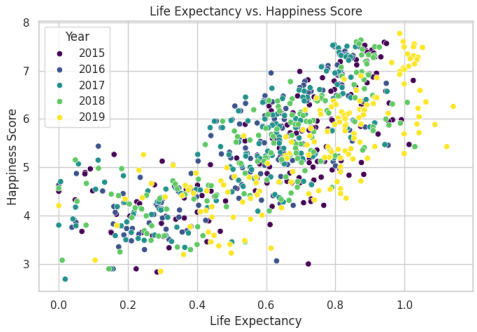
12.2 Scatterplots: Socio-economic Factors vs Happiness Score



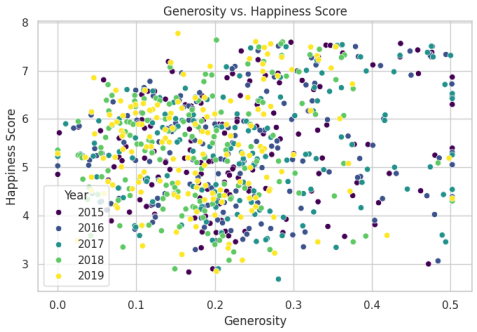
(a) Social Support vs Happiness Score



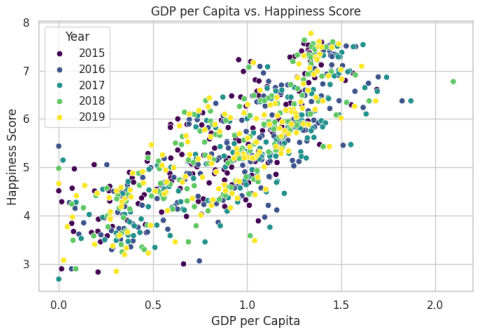
(b) Perceptions of Corruption vs Happiness Score



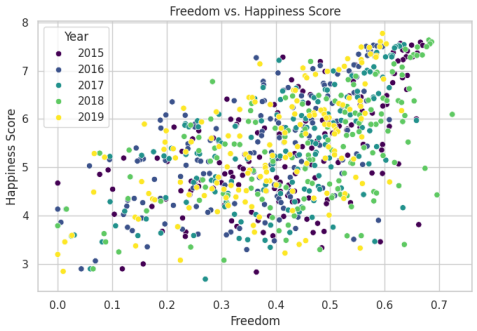
(c) Life Expectancy vs Happiness Score



(d) Generosity vs Happiness Score



(e) GDP per Capita vs Happiness Score



(f) Freedom vs Happiness Score

Figure 4: Scatterplots of Happiness Score against six socio-economic predictors (2015–2019). Each point represents a country-year observation; colour (if present in the image) denotes year. These plots visually confirm the positive relationships between Happiness Score and variables such as GDP per Capita, Life Expectancy, and Social Support, while showing weaker patterns for Generosity and Perceptions of Corruption.

Interpretation: The scatterplot grid in Figure 4 provides a comprehensive view of how each socio-economic factor relates to the Happiness Score. The key observations are below:

1. **GDP per Capita vs Happiness Score**
   A clear strong upward trend is visible. Countries with higher economic output per person tend to report higher happiness scores. The points form a tight rising diagonal pattern, confirming a highly linear relationship. This aligns with the strong correlation value ($r = 0.79$), indicating that economic well-being is one of the most influential predictors of national happiness.

2. **Life Expectancy vs Happiness Score**
   The scatterplot shows a strong positive association. Countries with higher life expectancy almost always exhibit higher happiness levels. The data points follow a consistent upward pattern, reflecting the high correlation ($r = 0.74$). This supports the idea that health and longevity play a central role in societal well-being.

3. **Social Support vs Happiness Score**
   A noticeable positive relationship is observed. Countries with stronger social networks—such as family, friends, and community support—tend to have higher happiness scores. Although slightly more spread out than GDP, the upward trend is clear and matches the correlation value ($r = 0.65$). This highlights the importance of emotional and social stability in influencing happiness.

4. **Freedom to Make Life Choices vs Happiness Score**
   The trend is moderately positive. While countries with greater personal freedom generally report higher happiness, the scatter is wider compared to the top predictors. This aligns with the moderate correlation ($r = 0.55$), indicating that although freedom contributes to well-being, its strength varies across nations.

5. **Perceptions of Corruption vs Happiness Score**
   The relationship appears weak and more dispersed. A slight positive trend exists, but the points are widely scattered. This visually matches the lower correlation ($r = 0.38$). The pattern suggests that corruption alone does not drive happiness levels, but may interact with other factors such as governance, trust, and social systems.

6. **Generosity vs Happiness Score**
   This scatterplot shows the weakest relationship. The points are widely spread with no clear linear pattern, reflecting the very low correlation ($r = 0.15$). This indicates that generosity may influence individual well-being more than national-level happiness indicators.

Overall Insight:
The scatterplots visually confirm that GDP per Capita, Life Expectancy, and Social Support have the strongest and most consistent positive associations with the Happiness Score. Freedom shows a moderate influence, while Perceptions of Corruption and

Generosity exhibit weaker and more dispersed patterns. These observations guided the selection of variables for the regression modelling phase, ensuring that only meaningful predictors were included.

12.3 Pairplot of Key Variables

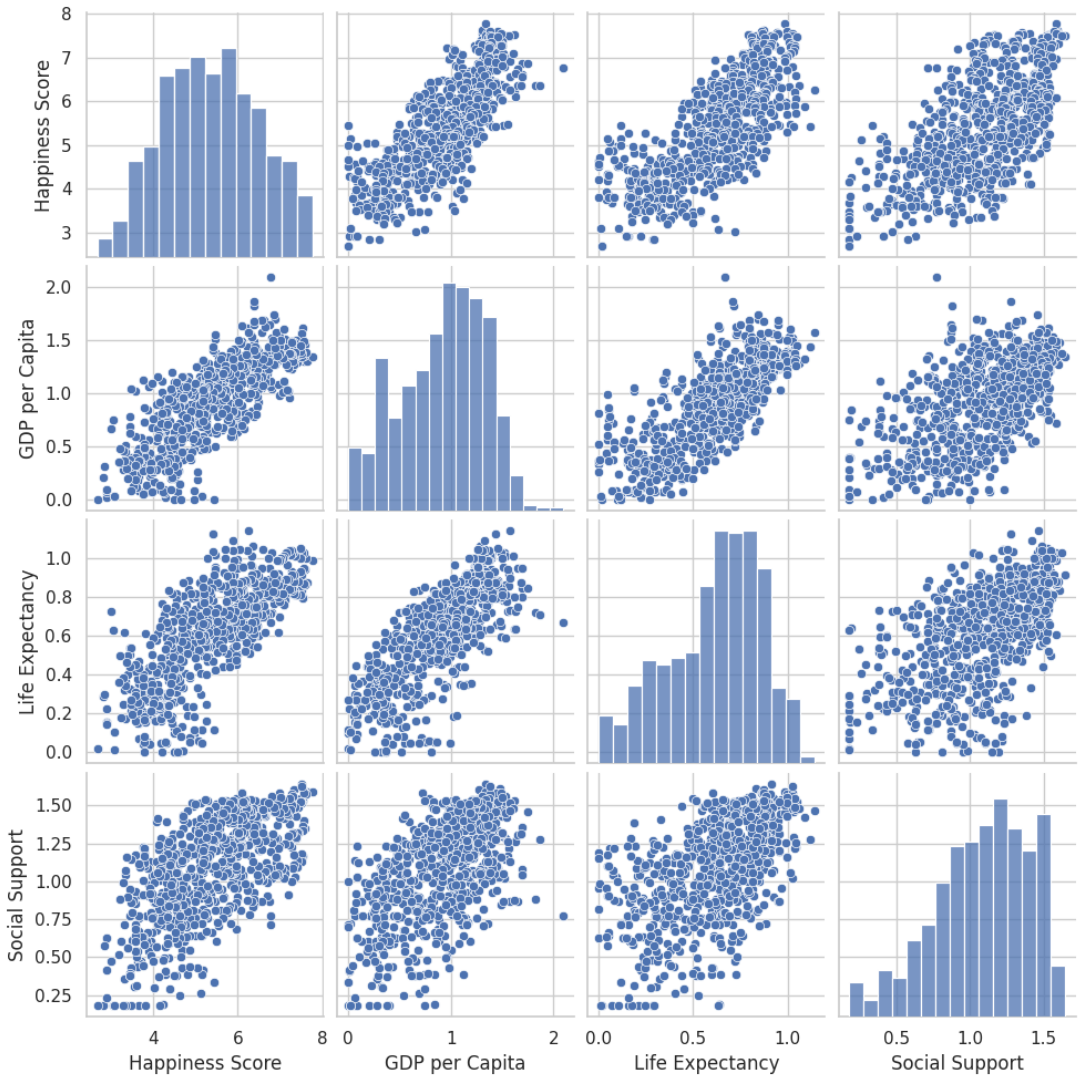


Figure 5: Pairplot of the four key variables: Happiness Score, GDP per Capita, Life Expectancy and Social Support. Diagonal panels show marginal distributions; off-diagonal panels show bivariate scatterplots and indicate strong linear relationships among the selected predictors.

Interpretation: The pairplot shown in Figure 5 provides a multi-dimensional view of how the key variables—Happiness Score, GDP per Capita, Life Expectancy, and Social Support—interact with one another. The plot serves as a compact summary of both the individual distributions and the bivariate relationships among these variables. The major

observations are detailed below:

1. **Diagonal Plots: Distribution of Each Variable**
   The diagonal elements of the pairplot display the histograms for each variable. These plots highlight how the data points are distributed:

   - *GDP per Capita* shows a slight right-skew, indicating that a small number of countries have very high economic output, while most countries lie in the lower-to-mid income range.
   - *Life Expectancy* and *Social Support* exhibit more symmetric distributions, suggesting a more even spread of values across countries.
   - *Happiness Score* shows a moderately concentrated distribution around its central range (between 4 and 7), reflecting similar global patterns in well-being.

   These histograms help us understand whether the variables are normally distributed, skewed, or contain clusters—information that is important for choosing appropriate modelling techniques.

2. **GDP per Capita vs Other Variables**
   The scatterplots involving GDP per Capita (first row/column off-diagonal) show clear upward trends with each of the other variables. Higher GDP per Capita is associated with higher Life Expectancy, stronger Social Support, and higher Happiness Scores. The linear structure of these plots visually confirms GDP's importance as a major socio-economic predictor.

3. **Life Expectancy vs Other Variables**
   Life Expectancy also shows strong positive relationships with both Happiness Score and Social Support. Countries with longer average lifespans tend to be happier and report stronger social connections. These scatterplots reinforce the numerical correlations observed in the heatmap.

4. **Social Support vs Other Variables**
   Social Support exhibits a strong positive correlation with both Happiness Score and Life Expectancy. Countries where people feel supported by their community not only show higher well-being but also tend to have better health outcomes. This suggests a close link between social cohesion and national development indicators.

5. **Happiness Score vs All Predictors**
   The final row/column shows the direct pairwise relationships with Happiness Score. All three predictors—GDP per Capita, Life Expectancy, and Social Support—display clear upward-sloping scatter patterns. This confirms that happier countries tend to be wealthier, healthier, and more socially connected.

Insight:
The pairplot provides strong visual evidence of the interconnected nature of key socio-economic indicators. The consistent upward trends between each predictor and the Happiness Score justify why these variables were selected as primary inputs for the regression analysis. The pairplot also highlights the strong interrelationships among the predictors themselves, reinforcing the idea that wealth, health, and community support collectively shape national happiness.

13 Regression Modelling

- **Model Building:** A Multiple Linear Regression model will be built. The model takes the form:

  $$\text{Happiness\_Score} = \beta_0 + \beta_1 * \text{GDP} + \beta_2 * \text{Life\_Expectancy} + \cdots + \epsilon$$

- **Model Evaluation:**

  [label=∘, itemsep=2pt]

  - **On Training Data:** The model's coefficients and statistical significance (p-values) will be examined to understand the impact of each predictor.
  - **On Testing Data:** The model's performance will be evaluated using the following metrics:
    - **R-squared (R^2):** The proportion of variance in the happiness score explained by the model.
    - **Adjusted R-squared:** R^2 adjusted for the number of predictors.
    - **Root Mean Squared Error (RMSE):** The average magnitude of the prediction errors, in the units of the happiness score.

- **Statistical Significance:** A p-value of less than 0.05 for a predictor's coefficient will be considered statistically significant, indicating a reliable relationship with the happiness score.

14 Regression Results

A multiple linear regression model was trained using time-series cross-validation to account for temporal dependencies in the data.The model was trained on the complete dataset and evaluated using robust temporal validation techniques.The cross validation was implemented with 4 splits to ensure temporal validation

14.1 Cross Validation Results

The model demonstrated consistent and improving performance across temporal folds:

```
Fold - R²: 0.7026, MAE: 0.4918, RMSE: 0.6159
Fold - R²: 0.7189, MAE: 0.3978, RMSE: 0.4977
Fold - R²: 0.6567, MAE: 0.4445, RMSE: 0.5728
Fold - R²: 0.7854, MAE: 0.4427, RMSE: 0.5803
```

Overall Cross-validated R^2: 0.7159 ($\pm$ 0.0923)

The overall cross-validated R^2 of 0.7159 is the most important metric. It indicates that, on average, our model can explain about 71.6% of the variance in the happiness scores of unseen data from a future year. This confirms that the model is not just memorizing the training data but has learned generalizable patterns.

15 Final Model Performance

The OLS regression model achieved strong explanatory power:

- R-squared: 0.768
- Adjusted R-squared: 0.765
- F-statistic: 340.6 (p-value: 2.28e-192)
- Number of Observations: 626

16 Statistical Significance & Model Equation

The regression analysis revealed highly statistically significant relationships for all predictors:

- **Model Equation:** Predicted Happiness Score = 2.1670 + 1.0056*GDP_per_Capita + 0.6613*Social_Support + 1.1522*Life_Expectancy + 1.3409*Freedom + 1.2325*Perceptions_of_Corruption + 0.5993*Generosity
- **Statistical Significance Analysis:**

  [label=○, itemsep=2pt]

  - **Highly Significant ($p < 0.001$):** All six predictors showed extremely high statistical significance.
  - **Strongest Predictors:** Freedom (coefficient: 1.3409), Perceptions of Corruption (coefficient: 1.2325), and Life Expectancy (coefficient: 1.1522) demonstrated the largest impact on happiness scores.

– **Notable Finding:** Unlike theoretical expectations, Perceptions of Corruption showed a positive relationship with happiness scores in this model, which warrants further investigation.

16.1 Comparison with Final Model

It is important to contrast the cross-validated score with the final model's R^2 of 0.768:

[leftmargin=*, itemsep=4pt]

- The final model's higher R^2 is expected because it was trained on the entire available dataset (2015-2019), maximizing its learning.
- The cross-validated score (0.716) is a more honest, pessimistic, and realistic estimate of how the model will perform when predicting a truly new, unseen year of data.
- The fact that these two numbers are relatively close (0.768 vs. 0.716) is a good sign, indicating that the final model is not severely overfitted.

In summary, the cross-validation provides strong evidence that our regression model is **reliable and generalizable**, capable of explaining approximately 71-77% of the variation in national happiness scores based on socio-economic factors, with a predictable and acceptable margin of error.

17 Final Conclusions

This project successfully developed a robust regression model that demonstrates a strong, quantifiable link between socio-economic factors and national happiness scores.

17.1 Project Synthesis: A 4-Step Methodology

Our team's analysis was conducted in four distinct phases, with each member building upon the previous one's work.

1. **Data Merging:** The project began by consolidating five separate annual CSV files (2015-2019) into a single, unified panel dataset of 782 entries. This crucial first step involved standardizing all column names to ensure consistency.
2. **Data Cleaning:** The merged dataset was then cleaned. The Interquartile Range (IQR) method was applied to detect statistical outliers. To avoid data loss, these outliers were handled using **capping** (setting values to the 1.5*IQR bound), which preserved the data points while neutralizing their power to skew the model.

3. **Exploratory Data Analysis:** With a clean dataset, a visual analysis was performed. The correlation heatmap and pairplots immediately identified the three dominant predictors: GDP per Capita ($r = 0.79$), Life Expectancy ($r = 0.74$), and Social Support ($r = 0.65$).

4. **Regression Modeling:** Finally, an Ordinary Least Squares (OLS) regression model was built to statistically quantify and validate these relationships.

17.2 Key Findings and Statistical Significance

The OLS model provided the definitive statistical proof for our project. The two primary indicators of the model's success are its R-squared and its p-values.

Table 1: Final OLS Model Performance

Metric	Result
R-squared (R^2)	**0.768**
Adj. R-squared	0.765
Cross-Validation R^2	0.716
p-values for all 6 predictors	< 0.05

An R^2 of **0.768** is the central finding. It signifies that **76.8% of the variance** in global happiness scores can be statistically explained by the six socio-economic factors in our model. The proximity of this value to the time-series cross-validation R^2 (0.716) confirms that our model is not overfitted and generalizes well to new data.

Furthermore, all six predictors were found to be *statistically significant* ($p < 0.05$), proving that their relationship with happiness is not a result of random chance.

17.3 Final Conclusion

Our project successfully met all stated objectives. We have demonstrated a clear, measurable, and statistically robust relationship between socio-economic factors and national happiness.

The findings from both the exploratory visualization and the regression model are in complete agreement: a nation's well-being is not abstract. It is powerfully and predictably linked to its **economic stability (GDP)**, **public health (Life Expectancy)**, and **social cohesion (Social Support)**. The key takeaway for any policymaker is that national happiness can be directly and reliably improved through investment in these core areas.