Analysis of Socio-Economic Factors on World Happiness

Project Report (2015–2019)

Aayush Randeep Deep Jangle Goutham Deepak Sarthak Gaur
Shrutisagar Jena

November 15, 2025

Project Objectives

1. **Merge Dataset:** Obtained data set have to be merged and the missing values have to be filled.
2. **Clean Data:** Detect and handle outliers using robust methods.
3. **Explore & Visualize:** Conduct EDA to uncover correlations between socio-economic factors and happiness.
4. **Model & Evaluate:** Build and validate a regression model to predict Happiness Score.

Merging 5 Years of World Happiness Data (2015–2019)

- **Load CSV Files** for each year via `pd.read_csv`.
- **Standardize Column Names** (e.g., *Economy (GDP per Capita)* $\rightarrow$ *GDP per Capita*).
- **Select Common Columns:** Country, Rank, Score, GDP, Social Support, Life Expectancy, Freedom, Corruption, Generosity.
- **Add a Year Column** for panel structure.
- **Concatenate All Years** using `pd.concat(..., ignore_index=True)`.
- **Clean the Data:** impute missing values, ensure numeric types, drop duplicates.
- **Export** as `cleaned_happiness_data.csv` (790 rows).

Result: Clean, unified 5-year panel dataset for analysis.

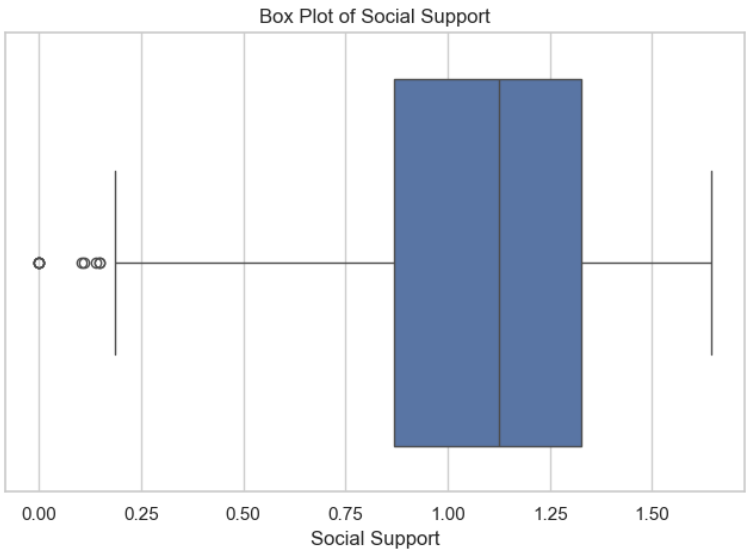
Outlier Detection

- After getting the merged data, it is now required to clean the data for further analysis.
- For good prediction, it is important that we remove the outliers.
- We have used the IQR Method to remove the outliers.

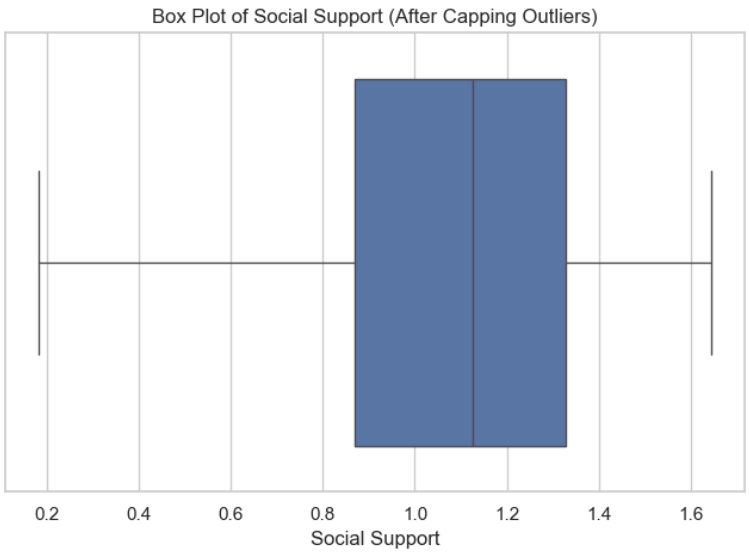
$$IQR = Q3 - Q1$$
$$\text{Lower Limit} = Q1 - 1.5\ IQR$$
$$\text{Upper Limit} = Q3 + 1.5\ IQR$$

- Data points lying below the lower limit and lying above the upper limit are called outliers.

Plots



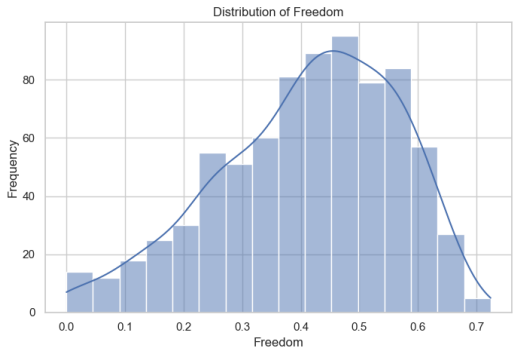
(a) Box Plot before Outlier Removal



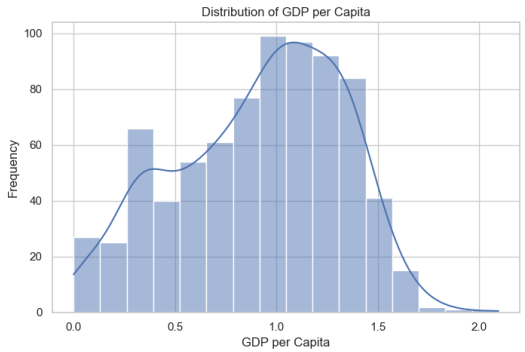
(b) Box Plot after Outlier Removal

For data consistency we use the *pandas.dataframe.clip* function which makes all the lower outlier points as the lower limit set by IQR and similarly for the higher side.

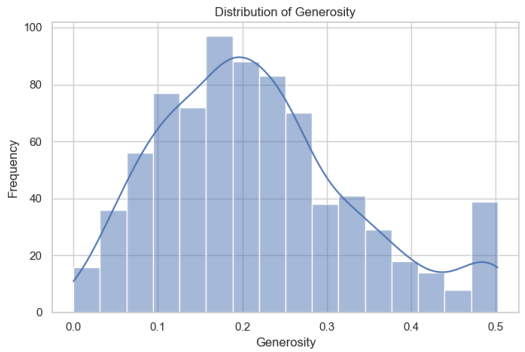
Distribution of data after Cleaning



(a) Freedom

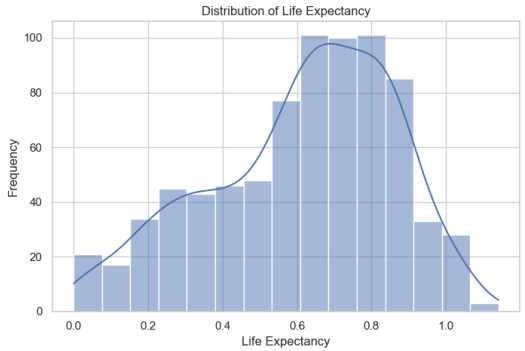


(b) GDP per Capita

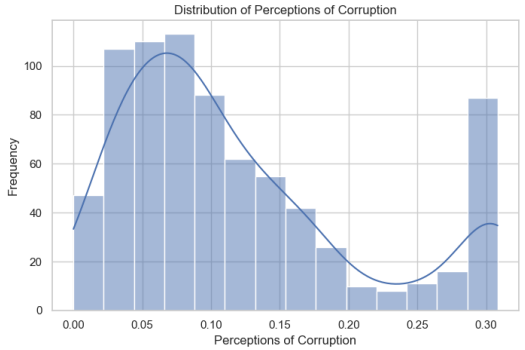


(c) Generosity

(d) Happiness Score

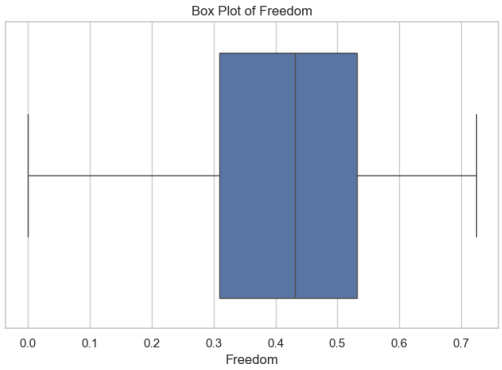


(e) Life Expectancy

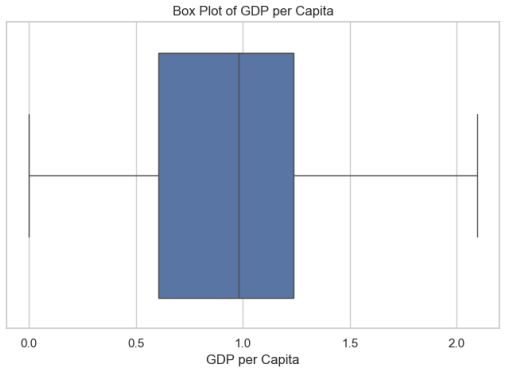


(f) Perceptions of Corruption

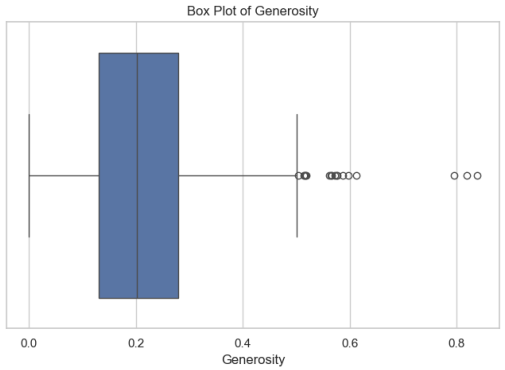
Box Plots After Cleaning



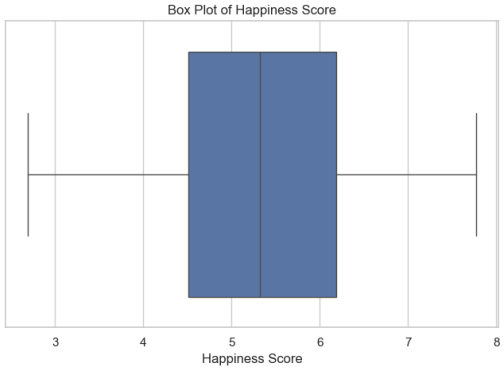
(a) Freedom



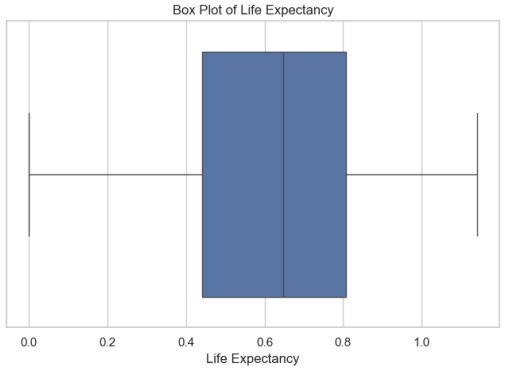
(b) GDP per Capita



(c) Generosity



(d) Happiness Score



(e) Life Expectancy

(f) Perceptions of Corruption

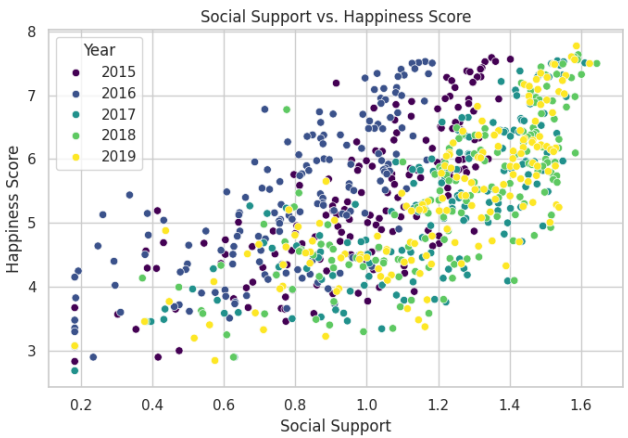
Correlation Analysis: Heatmap

- First step of EDA: identify strongest linear relationships with Happiness Score.
- Heatmap highlights dominant predictors visually.

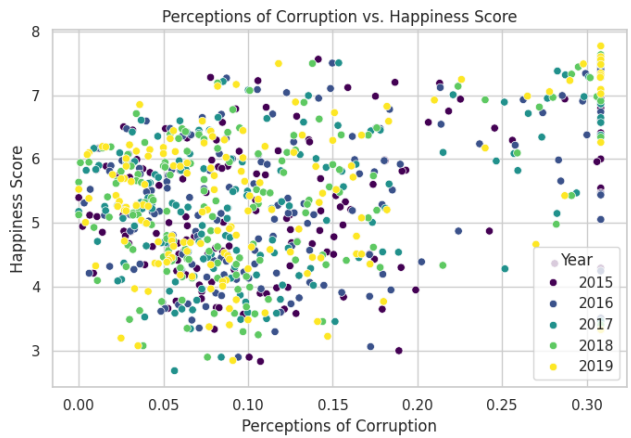
Correlation Heatmap: Key Insights

- The heatmap shows how socio-economic variables relate to the **Happiness Score**.
- **Strong predictors (high positive correlation):**
  - GDP per Capita ($r = 0.79$) — richer countries tend to be happier.
  - Life Expectancy ($r = 0.74$) — longer, healthier lives increase happiness.
  - Social Support ($r = 0.65$) — stronger community networks boost well-being.
- **Moderate predictors:**
  - Freedom ($r = 0.55$) — personal freedom improves life satisfaction.
  - Corruption Perception ($r = 0.38$) — lower corruption slightly increases trust and happiness.
- **Weak predictor:**
  - Generosity ($r = 0.15$) — has minimal effect at the national level.
- **Takeaway:** Heatmap helps identify meaningful predictors and check multicollinearity.

Scatterplots (1/3)

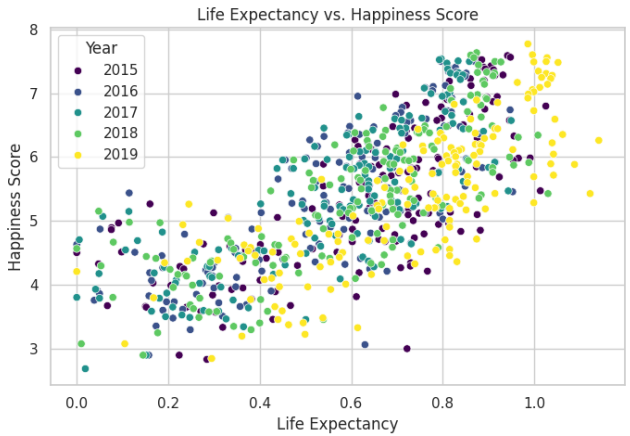


Social Support vs Happiness

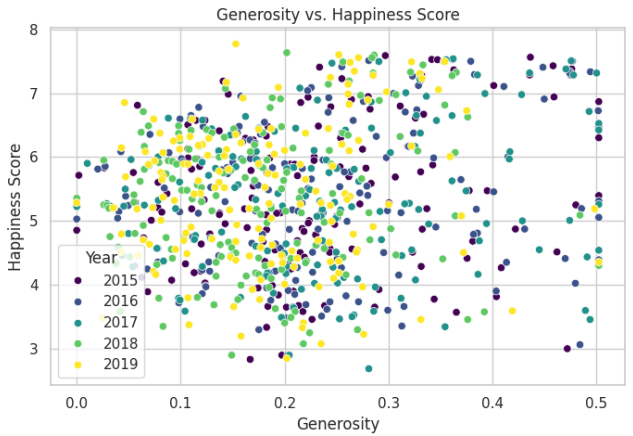


Perceptions of Corruption vs Happiness

Scatterplots (2/3)

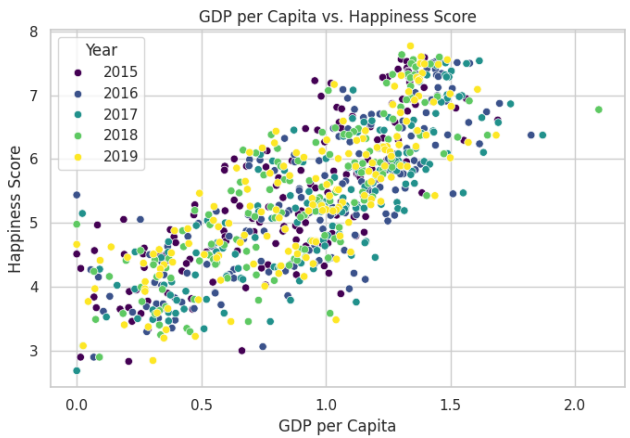


Life Expectancy vs Happiness

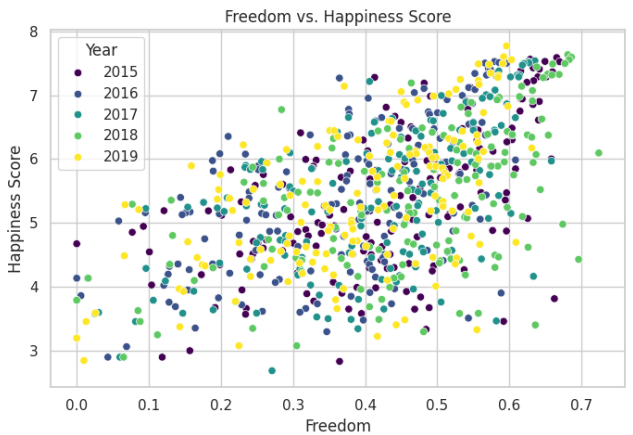


Generosity vs Happiness

Scatterplots (3/3)



GDP per Capita vs Happiness

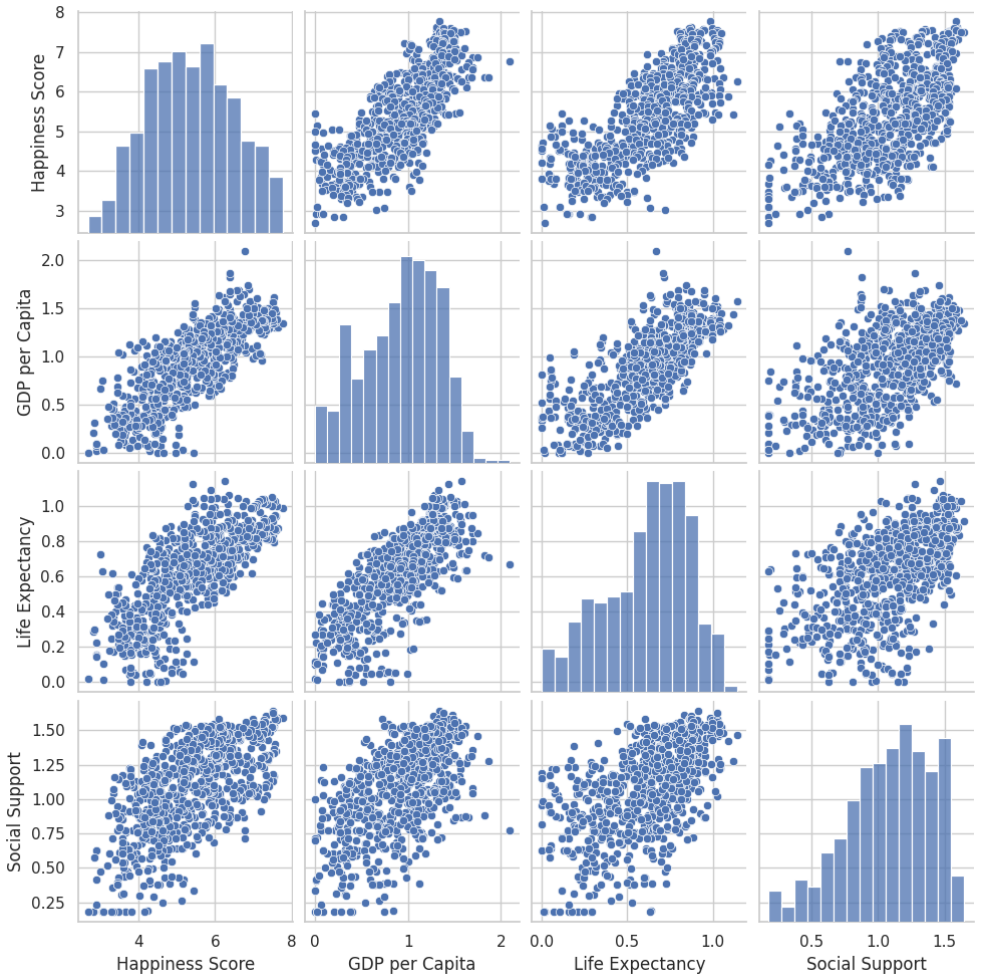


Freedom vs Happiness

Scatterplots: Key Interpretations (All Variables)

- **GDP per Capita ($r = 0.79$)** — strong upward trend; richer countries are consistently happier.
- **Life Expectancy ($r = 0.74$)** — longer, healthier lives strongly increase happiness.
- **Social Support ($r = 0.65$)** — stronger community/family networks lead to higher happiness.
- **Freedom ($r = 0.55$)** — moderately positive pattern; more personal freedom improves well-being.
- **Perceptions of Corruption ($r = 0.38$)** — weak positive trend; lower corruption slightly improves happiness.
- **Generosity ($r = 0.15$)** — weakest relationship; scattered points, minimal national-level impact.
- **Overall Insight:** Strong predictors = GDP, Life Expectancy, Social Support; Moderate = Freedom; Weak = Corruption, Generosity.

Pairplot of Key Variables



Pairplot: Key Interpretations

- **What it shows:** A multi-variable comparison of Happiness Score with GDP per Capita, Life Expectancy, and Social Support.
- **Diagonal (Histograms):**
  - GDP is right-skewed — few very rich countries.
  - Life Expectancy & Social Support are more symmetric.
  - Happiness Score clusters between 4–7 globally.
- **Off-Diagonal (Scatterplots):**
  - GDP shows clear upward trends with all variables — strongest predictor.
  - Life Expectancy rises consistently with Happiness and Social Support.
  - Social Support strongly aligns with both Happiness and Life Expectancy.
  - Happiness shows upward trends with all three predictors.
- **Overall Insight:** These variables rise together — wealth, health, and social cohesion jointly shape national happiness.

Regression Analysis of Happiness Scores

Regression Modelling Approach

- **Model Type:** Multiple Linear Regression
- **Model Form:**

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

Model Evaluation Strategy:

- Training Data: Analyze coefficients and statistical significance
- Testing Data: Evaluate predictive performance

Evaluation Metrics

- **R-squared (R^2):** Proportion of variance explained by the model
- **Adjusted R-squared:** R^2 adjusted for number of predictors
- **Root Mean Squared Error (RMSE):** Average prediction error magnitude

Statistical Significance Threshold:

- p-value ¡ 0.05 indicates reliable relationship
- Statistically significant predictors are retained

Cross-Validation Methodology

- **Technique:** Time-series cross-validation
- **Splits:** 4 temporal folds
- **Purpose:** Account for temporal dependencies
- **Goal:** Ensure robust temporal validation

Cross-Validation Results

Performance Across Folds:

- Fold 1: R^2: 0.7026, MAE: 0.4918, RMSE: 0.6159
- Fold 2: R^2: 0.7189, MAE: 0.3978, RMSE: 0.4977
- Fold 3: R^2: 0.6567, MAE: 0.4445, RMSE: 0.5728
- Fold 4: R^2: 0.7854, MAE: 0.4427, RMSE: 0.5803

- **Overall Cross-validated R^2:** 0.7159 ($\pm$ 0.0923)
- Explains 71.6% of variance in unseen data
- Model demonstrates generalizable patterns

Final Model Performance

OLS Regression Results:

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

- Strong explanatory power achieved
- All predictors statistically significant
- Model shows excellent fit to data

Model Equation

Final Regression Equation:

$$\begin{aligned}\text{Predicted Happiness Score} =& 2.1670 + 1.0056 \cdot \text{GDP\_per\_Capita} \\ &+ 0.6613 \cdot \text{Social\_Support} + 1.1522 \cdot \text{Life\_Expectancy} \\ &+ 1.3409 \cdot \text{Freedom} + 1.2325 \cdot \text{Perceptions\_of\_Corruption} \\ &+ 0.5993 \cdot \text{Generosity}\end{aligned}$$

Statistical Significance Analysis

- **All predictors:** Highly significant (p ¡ 0.001)
- **Strongest predictors by impact:**
  - Freedom (coefficient: 1.3409)
  - Perceptions of Corruption (coefficient: 1.2325)
  - Life Expectancy (coefficient: 1.1522)

Notable Finding: Perceptions of Corruption shows positive relationship with happiness scores, contrary to theoretical expectations. Requires further investigation.

Model Validation Comparison

Performance Comparison:

- Final Model R^2: 0.768 (trained on full dataset 2015-2019)
- Cross-validated R^2: 0.716 (realistic estimate for new data)

- Close values indicate minimal overfitting
- Cross-validation provides honest performance estimate
- Model demonstrates strong generalizability

Key Conclusions

- **Robust model** successfully quantifies link between socio-economic factors and happiness
- **Strong predictive power** explaining 71-77% of happiness score variance
- **All predictors** show high statistical significance
- **Model generalizes well** to unseen data with minimal overfitting
- **Reliable tool** for understanding national happiness determinants

Our Project: From Data to Discovery

As we've seen, our team followed a rigorous 4-step process:

1. **Data Merging:** First, we consolidated 5 years of data into a single, unified dataset of 782 entries.
2. **Data Cleaning:** Next, we managed the data quality, using the IQR method to find and cap all outliers for a robust analysis.
3. **Visual EDA:** Then, we explored the data using heatmaps and pairplots to visually identify the key drivers of happiness: GDP, Life Expectancy, and Social Support.
4. **Regression:** Finally, we built an OLS regression model to statistically prove these connections.

Our Key Finding: A Statistically Significant Link

Our analysis successfully proves that national happiness is not random.

Final Model $R^2 = 0.768$

This is our most important number. It means **76.8% of the variance** in world happiness scores can be explained by the 6 factors we analyzed.

Statistical Significance (p-values)

The model showed that **all six features** (GDP, Life Expectancy, Social Support, Freedom, Corruption, and Generosity) were statistically significant ($p < 0.05$).

This proves their impact is real and not just random chance.

Final Conclusion & Key Takeaways

- **Objective Met:** We successfully built a robust model that proves a strong, measurable, and significant link between socio-economic factors and national happiness.
- **What Matters Most:** The EDA and regression both confirm that **Wealth (GDP)**, **Health (Life Expectancy)**, and **Community (Social Support)** are the three most powerful predictors.
- **The Big Picture:** The key takeaway is that national well-being is not abstract. It is directly tied to concrete policy and investment in economic growth, public health, and social programs.

Thank You

On behalf of the entire team, we would now be happy to answer your questions.