Tab 1

**The secret behind “Happiness” – Insights from global statistics**

| **Group No.** | 4 | |
| --- | --- | --- |
| **Group Members** | **Name** | **Student ID** |
| <<\*Leung Pak Hei\*>> | <<\*1155245252\*>> |
| <<\*Tsang Chiu Kin\*>> | <<\*1155241250\*>> |
| <<\*Chan Wai Po\*>> | <<\*1155250656\*>> |
| <<\*Lee Yat Chun\*>> | <<\*1155249872\*>> |
| <<\*Tsoi Chung Him\*>> | <<\*1155242356\*>> |

**1. Research Background**

Global research on happiness investigates all aspects of a society that affect the well-being of everyone. The main objective of the study is to identify key factors, comparing outcomes across countries, and customize suggested policy measures to make the world a better place not just by looking at the economic indicators. There are several advantages behind the research:

* **Moving Beyond GDP**

Economic indicators e.g. GDP and unemployment rates do not capture all dimensions of life. Measures of wellbeing provide a better understanding of a society.

* **Policy Suggestion**

Factors like social trust, mental health, inequality, and the quality of governance are strongly related to life evaluations. Policy priorities, including preventive investments in mental health and initiatives that strengthen social capital are depending on those factors.

* **Effective of Social development**

Total happiness score is consistently associated with improved health, creativity, and labor market outcomes, All those factors should be taken into consideration for strategic and long-term sustainable development of a society.

**2. Research Questions**

1. Identify the most robust model for predicting happiness
2. Determine the strongest factor influencing happiness
3. Assess whether any single variable can sufficiently account for variations in happiness, or if a multifactor approach is required.

**3.1 Data Description**

The analytical dataset is constructed from the World Happiness Reports published by the UN Sustainable Development Solutions Network, using data sourced from the Gallup World Poll and distributed via Kaggle. After integrating the available annual files, the final dataset consists of 617 country–year observations, covering 167 countries across the years 2015, 2016, 2018, and 2019. These four years are the only ones with complete and internally consistent Gallup-based indicators. The 2017 dataset lacks the necessary variable structure and is therefore excluded.

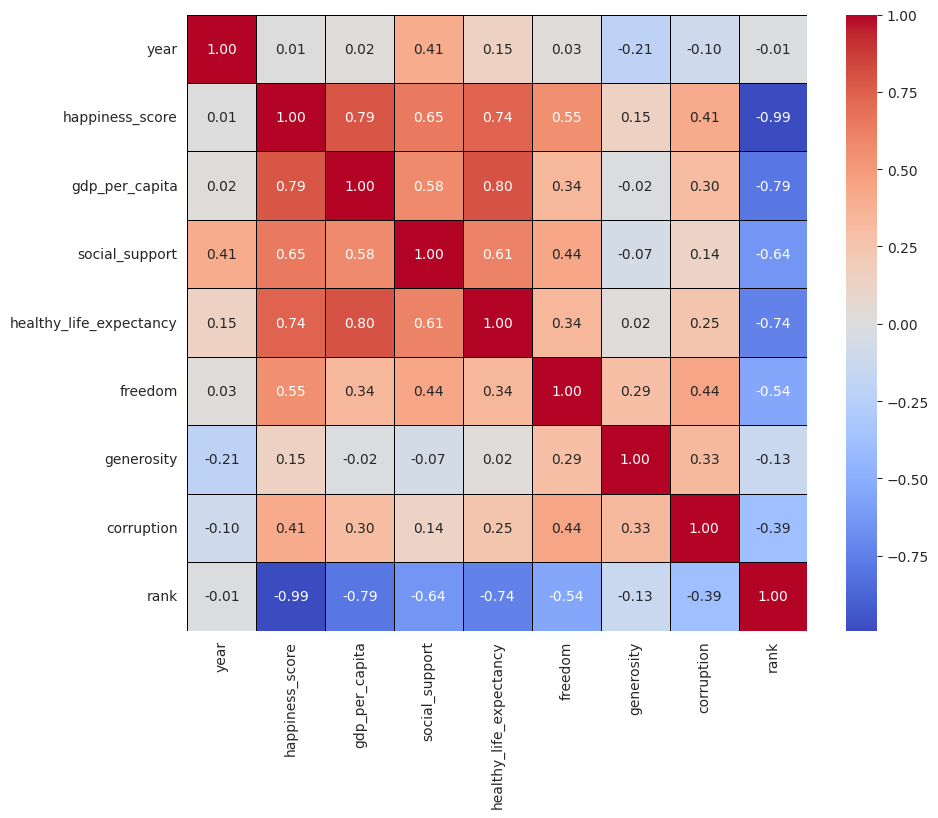
Each observation contains the happiness\_score and a consistent set of explanatory variables: gdp\_per\_capita, social\_support, healthy\_life\_expectancy, freedom, generosity, and perceptions\_of\_corruption, along with rank and region fields where provided. This unified schema enables direct comparison across countries and supports panel-style analysis of the structural, social, and institutional determinants of happiness.

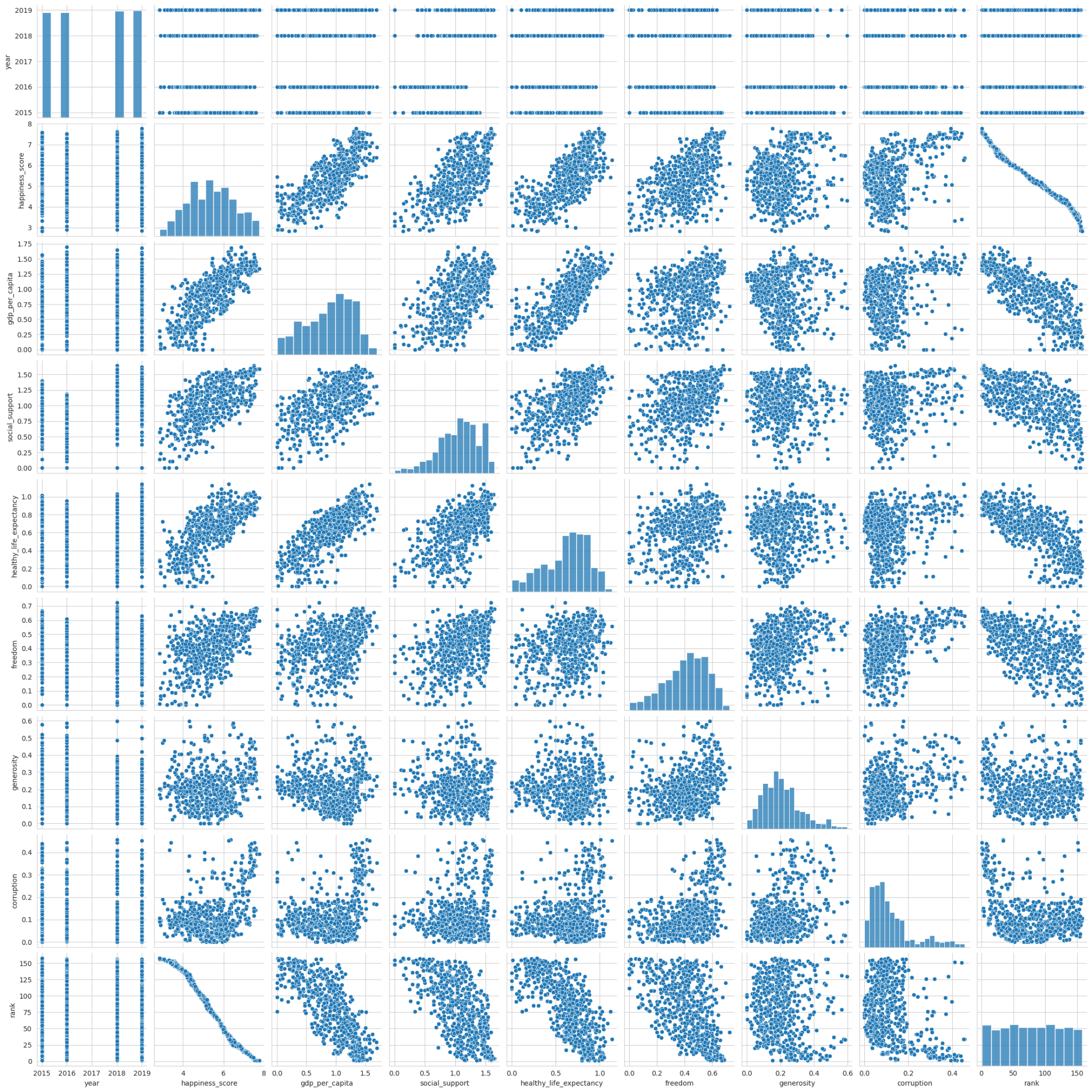
**3.2 Data Cleaning Process**

The cleaning workflow began by loading the five source CSV files and standardizing all column names into a uniform schema. Country names were reconciled to a consistent set of identifiers, after which all files were merged into a single long-format panel indexed by country and year. Observations with missing happiness scores or incomplete predictor sets were dropped, reducing 782 raw records to 617 complete entries, and duplicate country–year pairs were confirmed to be absent.

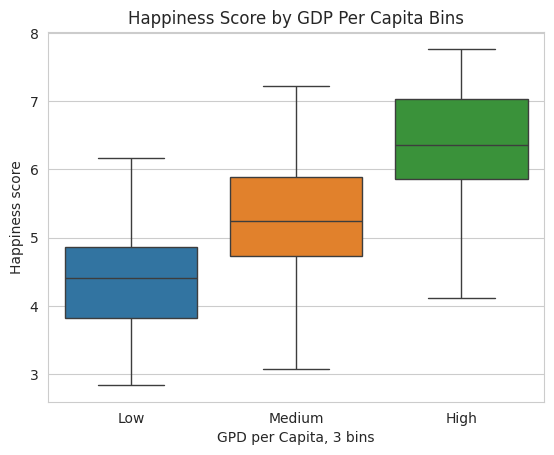
All numeric variables were cast to floating-point types and validated for reasonable ranges. Sparse missingness in the predictors was handled using within-country forward or backward filling, supplemented by region-level medians when temporal neighbors were unavailable. To mitigate undue influence on model estimates, nine outlier observations were removed using an IQR-based rule. Final integrity checks verified the uniqueness of each (country, year) combination and confirmed the completeness of all required fields. The resulting cleaned dataset provides a consistent and reliable foundation for the exploratory and regression analyses that follow.

**4.1 Exploratory Data Analysis (Key Drivers of Happiness**)





Gdp\_per\_capita(0.78) and Healthy\_life\_expectancy(0.74) are strongly and consistently positively correlated with happiness\_score . social\_support (0.65) and freedom (0.55) are other significant positive correlates of happiness.  
  
GDP per capita and healthy life expectancy show a strong positive correlation with happiness scores. We can interpret that economic prosperity and health are crucial determinants of a country's overall happiness. Countries with higher levels in these areas tend to enjoy greater collective happiness

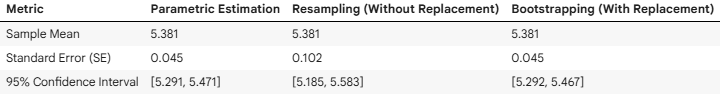


**4.2 Trend Over Time:**

For most countries, happiness\_score remained relatively stable between 2015 and 2019. However, a notable subset of countries exhibited high year-on-year variability fluctuations (std > 0.4) in their happiness scores highlighted in colors while all other countries remained stable over years. The year variable itself has a very weak correlation with overall happiness\_score (0.01).



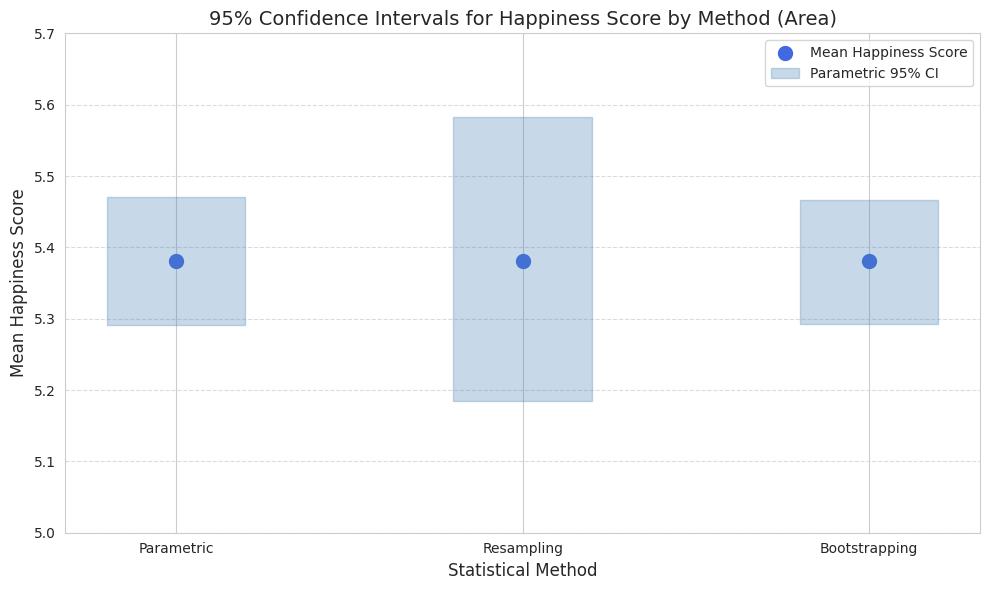
**4.3 Confidence in Mean Happiness Score**:



Three statistical estimation methods are used and they consistently estimated the mean happiness\_score to be approximately 5.381.

The Parametric and Bootstrapping methods produced nearly identical and narrower 95% Confidence Intervals ([5.291, 5.471] for Parametric and [5.292, 5.467] for Bootstrapping).  
  
The Resampling method, each resample used n = 100 resulted in a wider 95% Confidence Interval [5.185, 5.583] and a higher Standard Error (0.102), illustrating the impact of sample size on the precision of estimates.

To improve the precision and narrow the confidence interval for the Resampling method. We may consider increasing the sample size used in each resampling iteration.



**5.1 Regression Analysis**

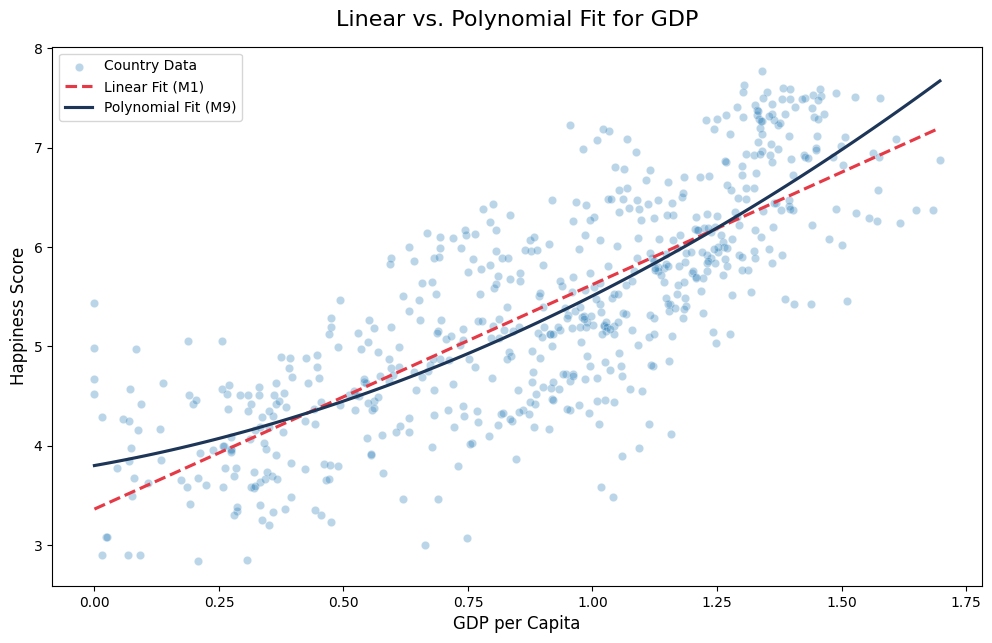
We used three regression models to identify the effect of the different predictors of happiness in different countries. First, we ran a simple linear regression model to look for individual relationships between each of the predictors and happiness level. Next, we used a multiple linear regression model to test the effects of all predictors and non-linear adjustments of "GDP per capita" (the predictor with the highest effect). Finally, we also ran a model with “Year” and “Country” as the fixed effects to identify their relationships with the six existing predictors.

**5.2 Simple Linear Regression (Models 1-6)**

In the baseline model, we ran linear regression models to study relationships between each of the individual predictors – GDP per capita, social support, life expectancy, freedom, generosity and corruption (Models 1-6). We found that the strongest predictor was GDP per capita, explaining 61.7% of the variance of happiness (Model 1).

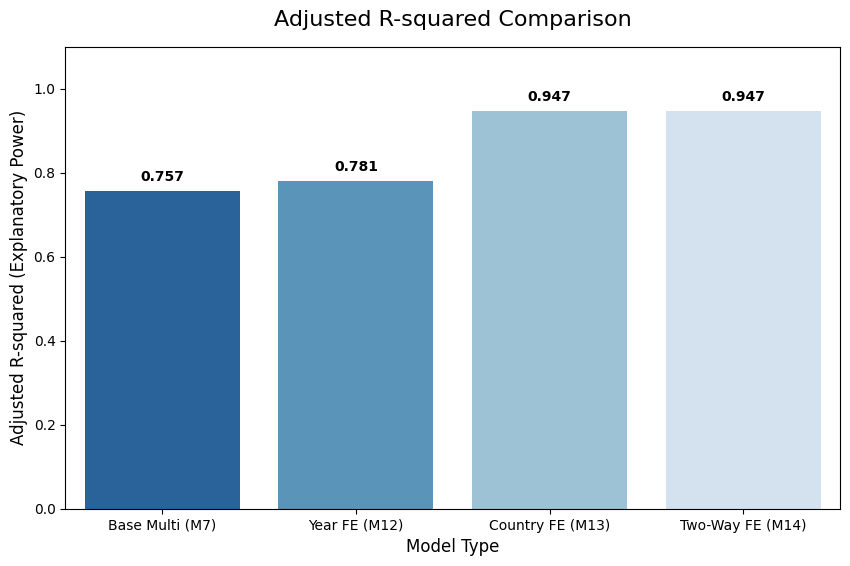
**5.3 Multiple Linear Regression (Models 7-11)**

We used a multiple linear regression model to study the effects of all predictors together on happiness scores in a multivariate model (Model 7). We also used (GDP per capita)2 and log (GDP per capita) and Log(Happiness Level) in different models as an additional variable (Models 8 to 11). The best model was Model 9 with (GDP per capita)2 outlining its non-linear relationship with happiness and all other predictors, which returned an Adjusted R2 of 76.0%.

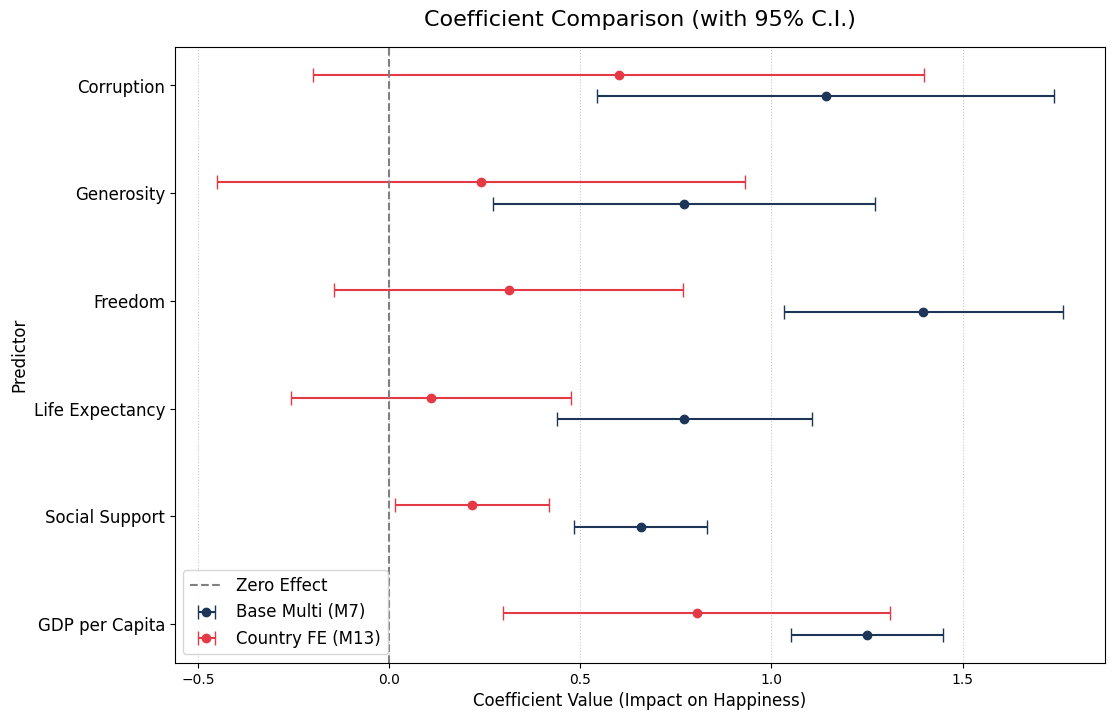


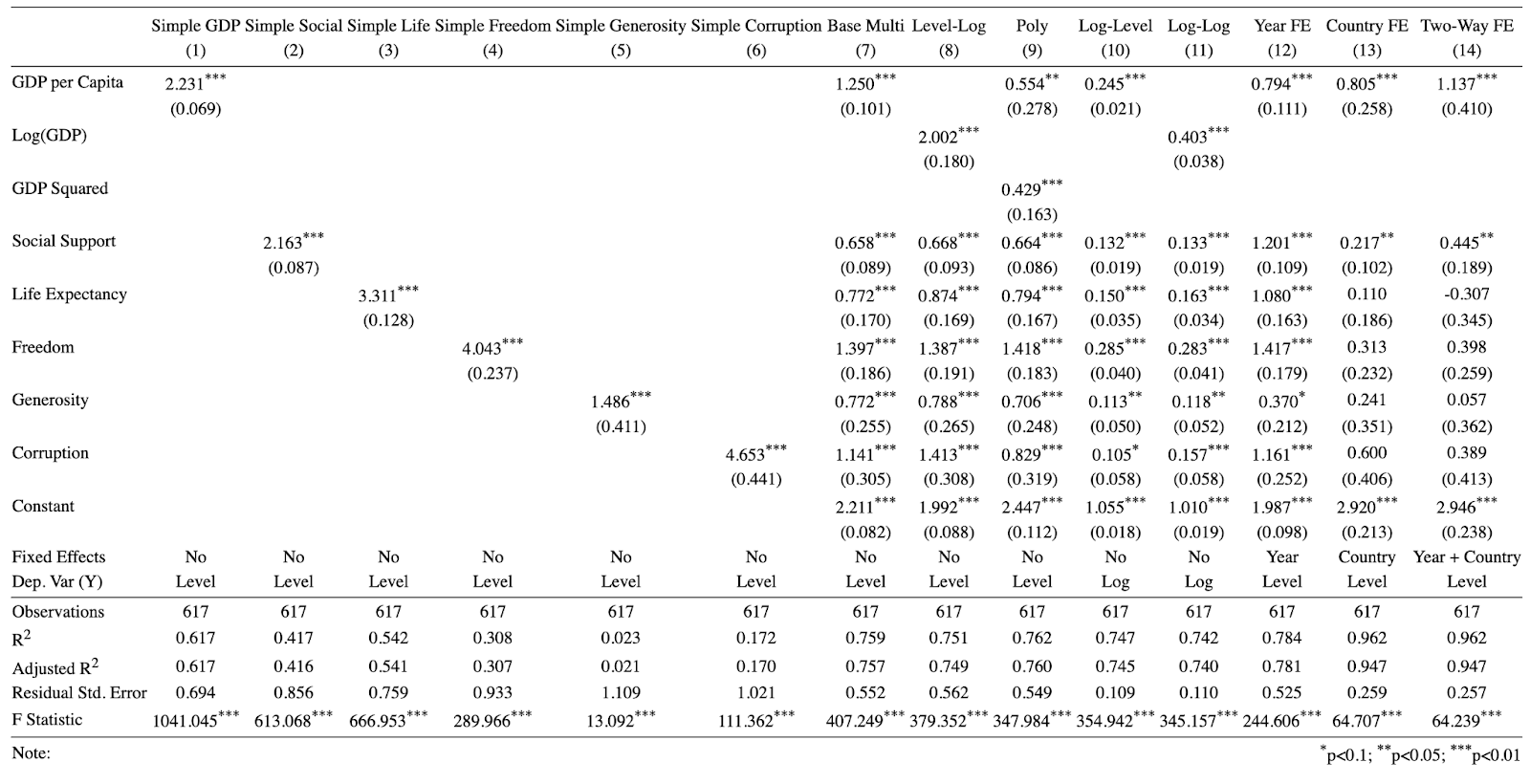
**5.4 Fixed Effects Model (Models 12-14)**

Using “Year” and “Country” as the fixed effects, we adopted a third model in an attempt to remove the noise of constant characteristics of individual countries and national events, e.g. an economic downturn, to reveal the real drivers of happiness scores. By incorporating the fixed effects, we controlled for both country-specific traits and specific years. The Adjusted R2 increased to 94.7% (Model 14).



After adding in the “Country” fixed effect of stable national traits, the data showed that four of the original predictors – Life Expectancy, Generosity, Freedom and Corruption are no longer statistically significant (p < 0.05). Only two predictors remain significant in this final model (M14) – GDP per capita (coeff: +1.137) and Social Support (coeff: +0.189). A possible explanation is that the positive results of these factors in the previous models are due to an omitted variable bias (OVB). Those are static features of those countries, rather than any independent drivers of change of happiness levels.



**6.1 Cross-Sectional Model Comparison (Complementary Analysis)**

While the above fixed effects models identified GDP per Capita and Social Support as drivers of country-specific changes, they cannot compare the relative importance of all six predictors across countries. To address this, we tested the following 8 candidate models (M1-M8) using train-test validation and standardized coefficients, enabling fair comparison of variables measured on different scales.

| **M1:** GDP only (baseline, 2 parameters) | **M5:** M4 + Generosity + Corruption (full linear model, 7 parameters) |
| --- | --- |
| **M2:** GDP + GDP² (testing nonlinearity, 3 parameters) | **M6:** M5 + GDP² (full with nonlinearity, 8 parameters) |
| **M3:** GDP + Health + Social Support (basic needs, 4 parameters) | **M7:** M5 + GDP×Corruption (interaction model, 8 parameters) |
| **M4:** M3 + Freedom (adding autonomy, 5 parameters) | **M8:** GDP + GDP² + Social + Health + GDP×Social (alternative interaction, 6 parameters) |

**8 Candidate Models Tested:** 70/30 train-test split with 5-fold cross-validation. Evaluation metric: Root Mean Squared Error (RMSE) on test set.

A comparison of a bar graph

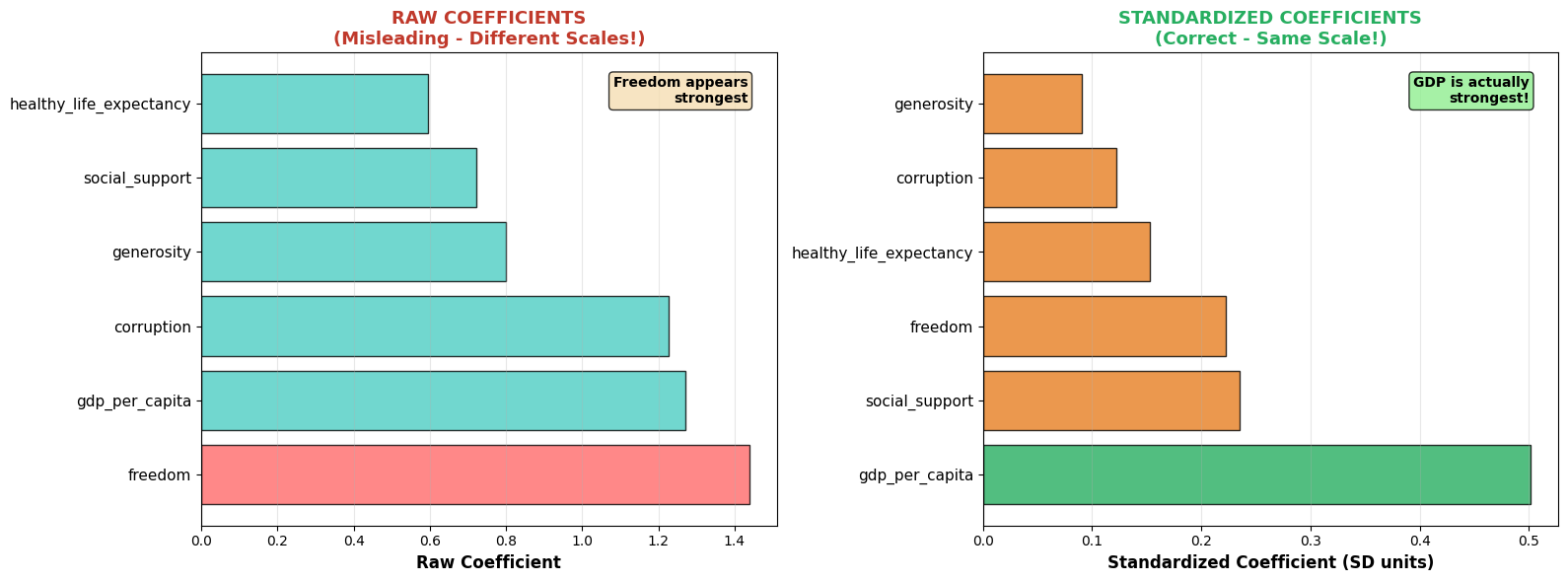
AI-generated content may be incorrect.

**Selection Result:** **Model 5** selected as optimal. Test RMSE: 0.5881 (only 0.0007 behind best model M7), Adj. R²: 76.3%, all 6 predictors statistically significant (p < 0.01). M5 achieves near-optimal performance with maximum simplicity, one fewer parameter than M7/M6 with no meaningful performance loss.

**6.2 Standardization reveals True Variable Importance**

Variables have vastly different measurement scales (GDP's SD = 0.395; Corruption's SD = 0.100). Comparing raw coefficients is misleading because a "1-unit change" means entirely different things for each variable.

**Raw vs. Standardized Rankings:**

****

**Key Insight:** Freedom appeared to have the largest coefficient (1.44) in raw terms because its small standard deviation (0.154) meant the model needed a large multiplier to produce meaningful effects. Once we account for this scale difference through standardization, we discover:

* **GDP's true effect (0.502) is 2.26× larger than Freedom's (0.222)**
* **GDP's true effect (0.502) is 4.08× larger than Corruption's (0.123)**

**Rankings Completely Reversed After Standardization:**

| **Variable** | **Raw Coef.** | **Standardized Coef.** | **True Rank** | **% of Total Effect** |
| --- | --- | --- | --- | --- |
| GDP per capita | 1.27 | **+0.502** | **1st** | **37.9%** |
| Social support | 0.72 | **+0.235** | **2nd** | **17.7%** |
| Freedom | 1.44 | **+0.222** | **3rd** | **16.8%** |
| Healthy life expectancy | 0.59 | **+0.153** | **4th** | **11.5%** |
| Corruption | 1.23 | **+0.123** | **5th** | **9.3%** |
| Generosity | 0.80 | **+0.091** | **6th** | **6.9%** |

**Critical Insight:** Without standardization, we would have concluded Freedom is most important—**completely wrong**. GDP per capita's true effect is **2.26× larger** than Freedom's (0.502 vs. 0.222).

**Real-World Validation: Qatar vs. Costa Rica**

| **Qatar:** GDP per capita $128K, Happiness Rank #58 | **Costa Rica:** GDP per capita $12K, Happiness Rank #12 |
| --- | --- |

Why is Costa Rica 46 ranks happier despite 10.7× lower GDP?

| **Factor** | **Qatar** | **Costa Rica** | **Difference** |
| --- | --- | --- | --- |
| Freedom | 0.60 | 0.95 | +0.35 (+58%) |
| Social support | 0.85 | 0.91 | +0.06 (+7%) |
| Corruption control | 0.40 | 0.77 | +0.37 (+93%) |

**Calculation: Using our Model 5:**

Qatar's GDP advantage: +$116K → +0.40 happiness boost

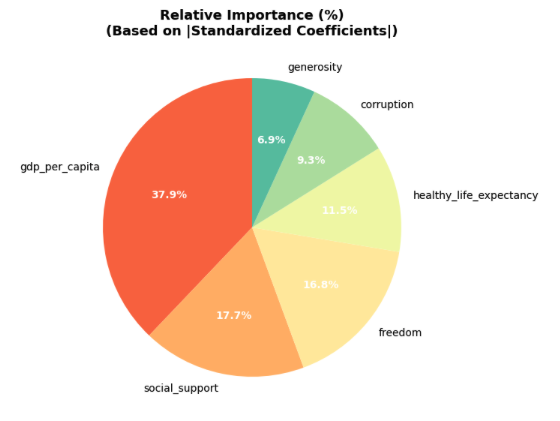
Costa Rica's advantage in other 5 factors → +0.65 happiness boost

Net effect: Costa Rica wins by + ~ 0.25 happiness points

The Lesson: Money cannot buy what good governance, strong communities, and personal freedom provide. Focusing solely on GDP captures only 38% of achievable gains. The other 62% comes from social, health, governance, and cultural dimensions.

**Methodological Clarification:** The fixed effects model 5.4 identifies determinants of within-country variation when Time-invariant factors are absorbed and unestimable. The cross-sectional model 6 identifies determinants of between-country differences, precisely those stable characteristics. Qatar-Costa Rica exemplifies how baseline happiness is multidimensional even when year-to-year changes depend primarily on GDP and social support.

**6.3 Policy Recommendations based on M5 Framework**

****

1. Economic Development (37.9%): Sustainable GDP growth and income equality
2. Social & Health Infrastructure (29.2%): Strengthen social safety nets, community programs, and healthcare access
3. Governance & Cultural Factors (32.9%): Enhance personal freedoms (16.8%), reduce corruption (9.3%), foster generosity (6.9%)

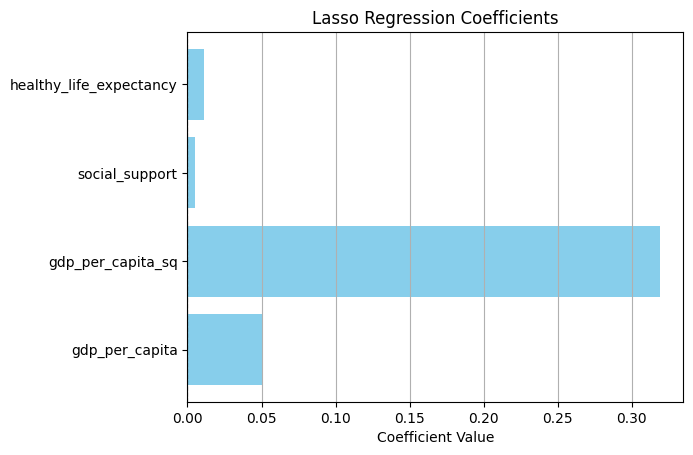
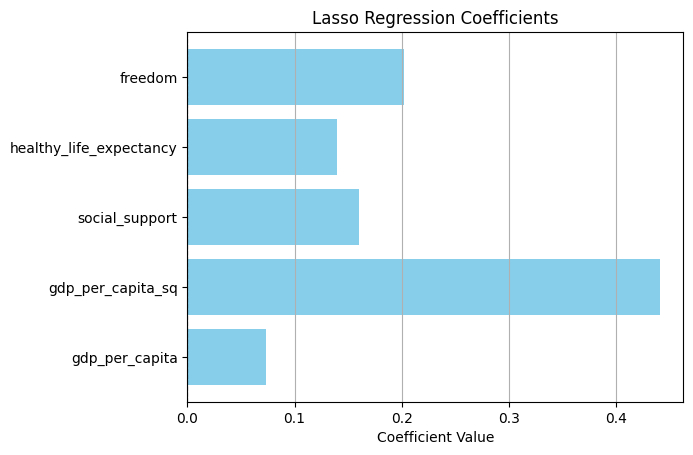
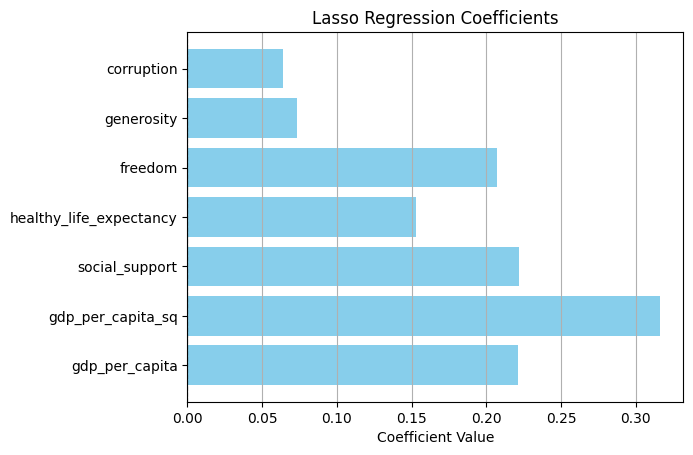
National happiness requires balanced development across all six dimensions and model 5 demonstrates excellent predictive power while maintaining interpretability. Standardization is critical from the lesson learnt, without it we would have drawn wrong conclusions about which factors matter most. In addition to model 5 as optimal for interpretation, we also need to test whether regularization methods could improve predictive accuracy.

**7 Regularization and Lasso Regression**

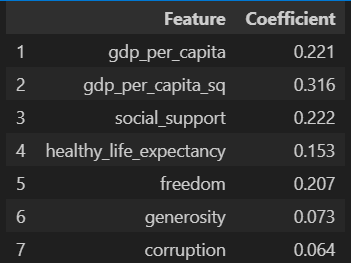
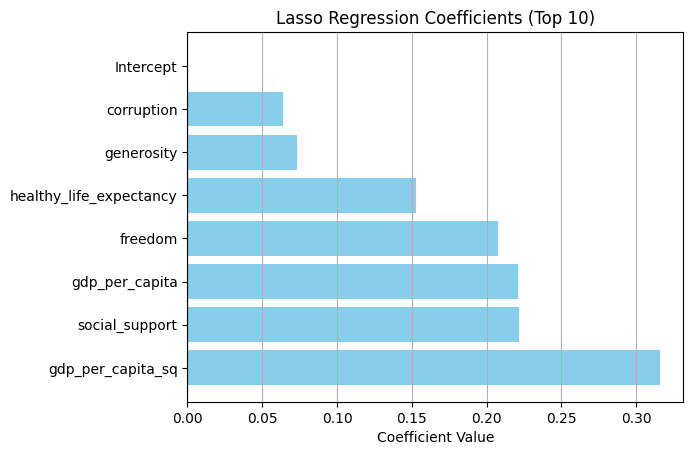
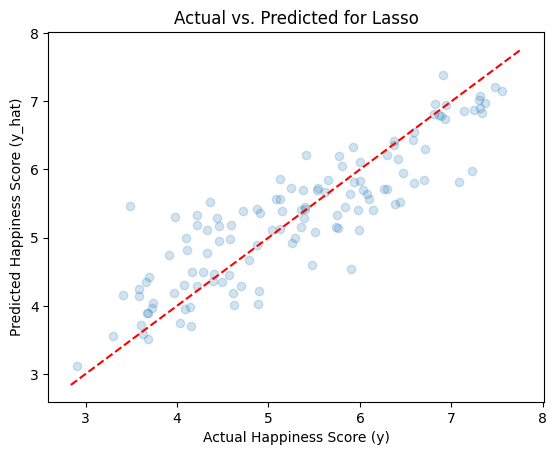
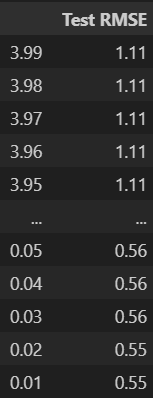
To maximize prediction accuracy beyond Model 5 above, we applied Lasso regression with alpha 0.01 to all 8 models.

| **Model** | **RMSE** | **Model** | **RMSE** |
| --- | --- | --- | --- |
| M1 | 0.6928 | M5 | 0.5495 |
| M2 | 0.6799 | M6 | 0.5460 |
| M3 | 0.6273 | M7 | 0.5490 |
| M4 | 0.5673 | M8 | 0.6074 |

Selecting Model 6, based on a 20% test set, we increase alpha from 0.01 to 0.1 to 0.5 (from top to bottom below), RMSE increase from 0.5532, to 0.5953, then to 0.9000. Lasso selects GDP per capita square as the most important dependent variable, whereas sacrificing explanatory power of the regression model.



After testing with a range of alpha from 0.01 to 4 on the 0.01 division scale, the optimal alpha should be 0.01 when RMSE is lowest. The most important independent variables found are GDP per capita square, social support, freedom and healthy life expectancy, with GDP per capita showing a non-linear relationship with the happiness score. It is believed when GDP per capita increases, the increase in happiness will increase at an even higher pace, particularly when a country emerges from developing to developed status. On the actual vs predicted chart, the prediction model basically can capture the trend and most of actual values, except few outliers.

**8 Conclusion**

Compared to the earlier result from cross validation, Lasso regression helps reduce the RMSE from 0.5891 to 0.5460 (for Model 6). It is found that Model 5 is useful for interpretation, whereas Model 6 is more useful for accurate prediction, and both models highlight the importance of GDP per capita. Based on the results from different models we studied, it is recommended for countries to improve on below policy areas to increase the national happiness level.

* Other than targeting GDP growth, the distribution of GDP and economic benefit to society is important.
* Providing education and training to students and workers to catch up with the pace of AI/technology investment by corporations, so that the growth from AI would not be concentrated by a small number of corporations, and more populations can contribute to the growth in the industry and thus the economy.
* More health care support, for example, public hospital service, or insurance policy like Obamacare in the US.
* Clean environment policy, for example electric vehicles, renewable energy, and enhancing public facilities for public health protection.
* Providing more freedom for the society to participate in the economic growth, policy direction and innovation.