Springboard DSC
Capstone Project 2

Modeling Feature Impact on World Happiness Data

Ross Brinkerhoff
May 2024

1. Introduction

This project consists of analyzing and modeling a dataset from the World Happiness Report published by the Sustainable Development Solutions Network.

The report, and this data set, is based on a collection of global survey data that measures the perceptions of the citizens in different countries across a variety of features. Some example include perceptions of corruption in government, freedom to control one life etc. All of these questions are combined with country GDP and life expectancy data to produce a broad picture of the situation in a particular country over time.

The key metric, and our dependent variable for this project, was the life ladder question. This question asks the respondent to rank their life as a whole from one to ten, with one being the worst possible life and ten being the best possible life.

The objective of the project was to obtain a deeper understanding of how the various features in the dataset affect the variability of citizen responses to the life ladder question. In other words, which features are most important in explaining a given life ladder response, and what impact do the individual features have on that response?>

The project was able to successfully model the data and identify the top five most important features based on that model. We were able to gain an understanding of the impact of each of those key features on the target variable and to identify thresholds where each feature starts to drive positive increases in the target.

Implementation details can be found below, as well as in the notebooks. Details on individual features can be found in the data dictionary.

2. Approach

2.1 Data Acquisition and Wrangling

A single main dataset was used for all modeling and analysis work.
The dataset is freely available on Kaggle.

Supplemental data available from the World Bank and Encyclopedia Britannica was also consulted for filling gaps in the data where necessary

In its original form, the data consisted of rows for 2,199 entries, each corresponding to a year and a particular country. In addition to the life ladder score/target variable, there were 12 other features in the data. Two of these features were categorical identifiers (Country Name and Region), one was temporal (Year), and two were based on outside data sources that had been incorporated into the data set (Log GDP and Healthy Life Expectancy). The remaining seven features consisted of float values corresponding the various survey question scores for a given country in a given year.

It's important to note that the dataset used in this project had already been processed by the publisher to some extent. It is aggregated responses based on survey data, not the survey data itself.

Additional information on what these survey questions refer to or how the values were arrived at can be found at the link above for the data set, or at the website for the ongoing Happiness Report Project

The main challenges in terms of data wrangling consisted of imputing or dropping missing values and determining the best strategy for dealing with outliers.

Missing values

As with most data sets, there were some missing values in the data. The early years of the data set (2005 and 2006) have significantly fewer entries than the other years (2007-2022), all of which were represented with between 104 and 147 rows.

For features without a lot of variability and with only occasional missing values (such as GDP), the missing values could just be imputed. They were sometimes also manually adjusted in the case of missing data for Regional indicator.

When it came to the survey data features, the approach was less clear. Many rows had multiple missing values for survey responses, and some had no values at all for some features. In particular, the "Confidence in National Government" feature was missing for several countries, as was, to a lesser extent, the "Perceptions of Corruption" feature. It was not clear how to impute these values, based as they were on survey data and given that they

were one of a relatively small number of features in the data set. In addition, in countries where the value was mostly or completely missing, there was no basis on which to impute a new value. The general approach to these issues was to drop rows with multiple missing values, or with large numbers of missing values in a particular feature. Imputation using the mean was used on countries with three or fewer missing values for the Confidence in National Government feature.

Dropping the rows using the above process cleared up almost all of the missing value issues in the data. The trade off was that it also resulted in some very significant countries being dropped and not being used in further modeling.

The remaining missing values were for life expectancy. These were filled in using data from the World Bank on life expectancy.

Outliers

With most of the data based on survey responses, it was unclear what a true outlier might be in many cases involving those features. The focus for outlier corrections was therefore on features like GDP and Healthy Life Expectancy, where some intuition about outliers was possible.

Examination of the data revealed a small number of outliers that we investigated in more detail, since they seemed to be beyond the realm of reasonable possibility. The two most outstanding outliers in this regard were a minimum value of 6.9 years for Life Expectancy and maximum value for Confidence in National Government of 0.99.

The minimum life expectancy value turned out to be from Haiti in 2006, and appeared to be related to multiple catastrophes in that country in the 2010s. An average life expectancy of 6.9 years did not seem credible.

An additional challenge arose at this stage, in that that the life expectancy data in the World Happiness data set did not appear to match the data provided by the world bank. This seems to be the result of World Happiness data using WHO data for some years, and extrapolating values for other years. Instead of substituting WHO value for all the life expectancy value in the Happiness data set, it was decided to preserve the methodology from the publishers and just drop the few rows from Haiti that appeared to be obvious outliers.

The high confidence in government outliers came predominantly from Rwanda, with scores indicating basically complete confidence in the government over multiple years. The highest score from Rwanda was from 2016, at which time the country appears to have had a repressive regime. There were a number of countries in a similar circumstance, and these were ultimately left as is since it was not clear what adjustment might be appropriate, if any.

After the data cleaning process described above, 348 rows had been dropped, leaving a data set consisting of 1,851 rows containing data from 18 different years, 139 different countries and 10 different regions.

2.2 Storytelling and Inferential Statistics

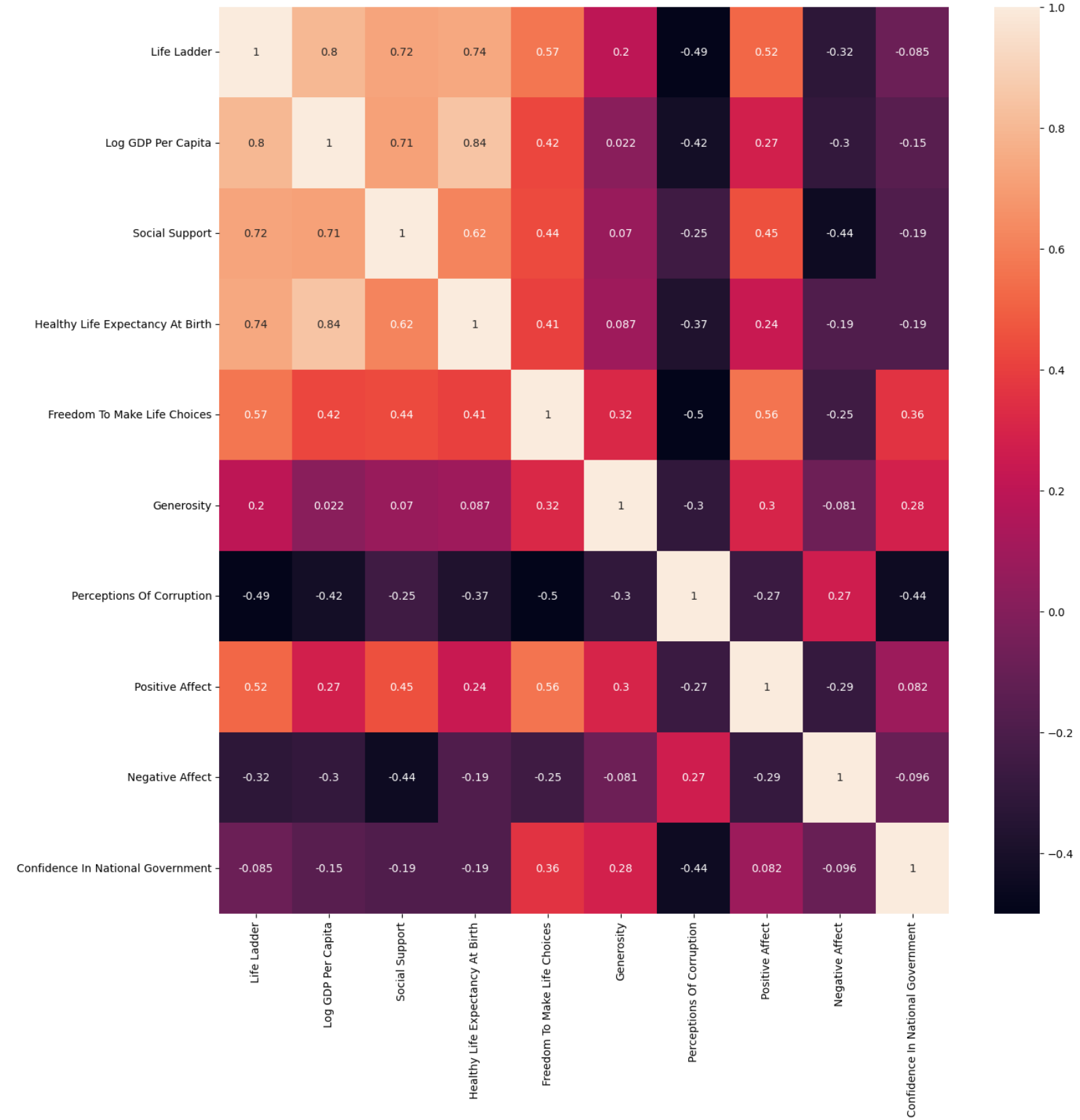
EDA work was focused on finding obvious correlations in the data, and on determining how much of the world was represented in the data after cleaning. Not every country was present in the data before cleaning, so there was a concern that dropping even more data during the cleaning process could skew the potion of the world represented still further. The goal of the project was to understand the relationships between features and the target variable on a global scale, so having a dataset skewed, for example, towards wealthier countries, would have been an issue.

Correlation

Using a heat map(shown below), several numeric features were identified as having high correlation with the target. These included, most notably, Log GDP, Healthy Life Expectancy, and Social Support, with less notable correlations for Freedom to Make Life Choices and Positive Affect. Perceptions of Corruption was the most negatively correlated with the target.

Correlations between features were most notable in Log GDP and Life Expectancy (0.84), with other features also showing some positive correlations. That being said, many features were not strongly correlated with each other or the target. The features that were most strongly correlated with the target served as the basis for further exploration in the modeling phase.

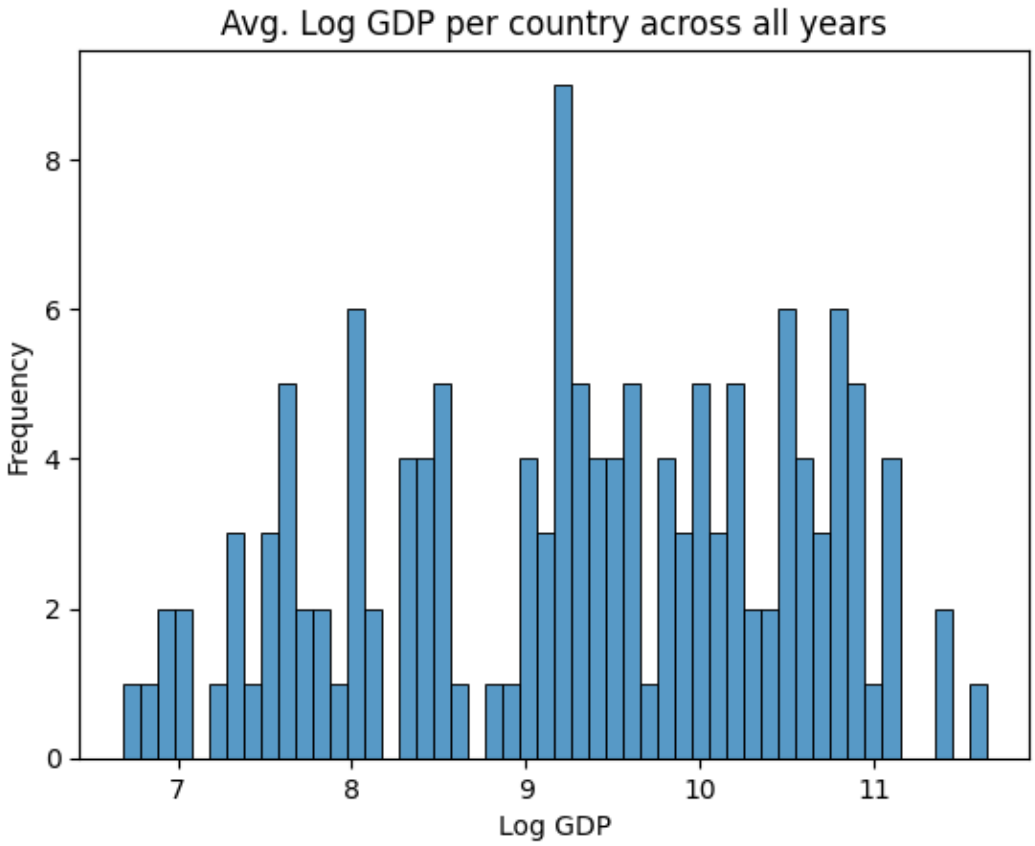
Heat Map of all Features



Economic Representation

The main concerns in terms of representation were economic and regional. It was important to determine if the data was significantly biased to towards wealthy countries or poor countries. It was also important to determine which regions, if any, were more represented.

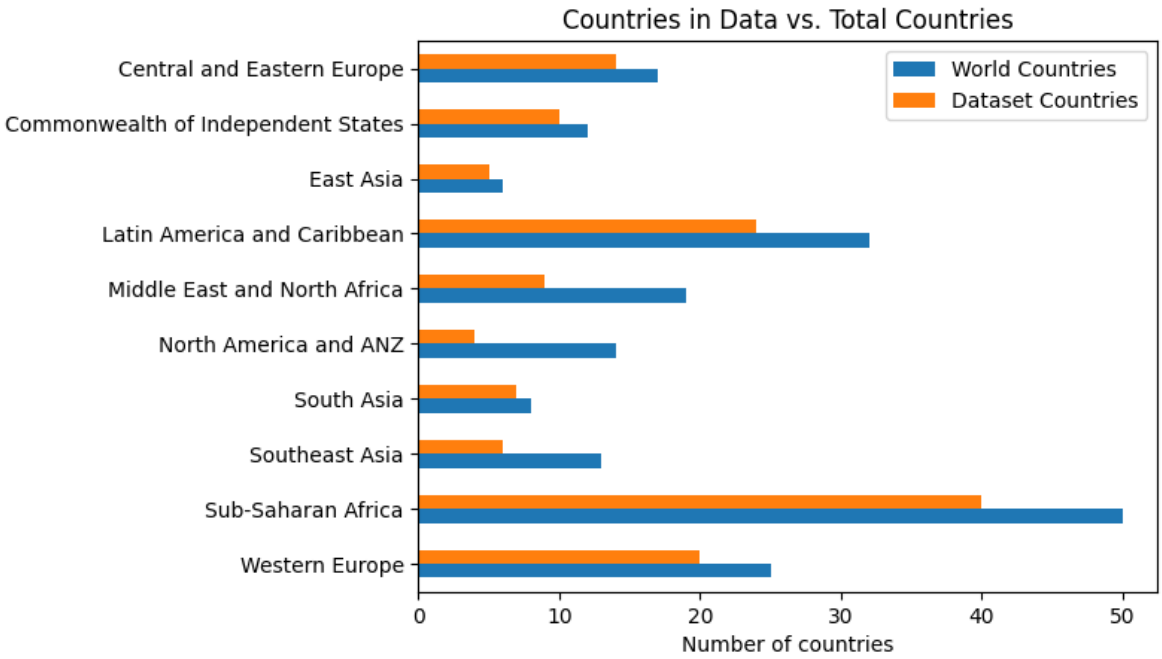
To explore the economic representation question, we looked that the distribution of the Log GDP feature, which was the only economic indicator in the dataset. Log GDP values did not vary significantly from year to year, so we computed and plotted the average Log GDP for each country across the data set. The distribution does not show a significant skew and has no outliers, indicating that the data available was not significantly biased towards either wealthy or poor countries.



Mean	Standard Deviation	3 SD above Mean	3 SD below Mean
9.3	1.22	12.96	5.64

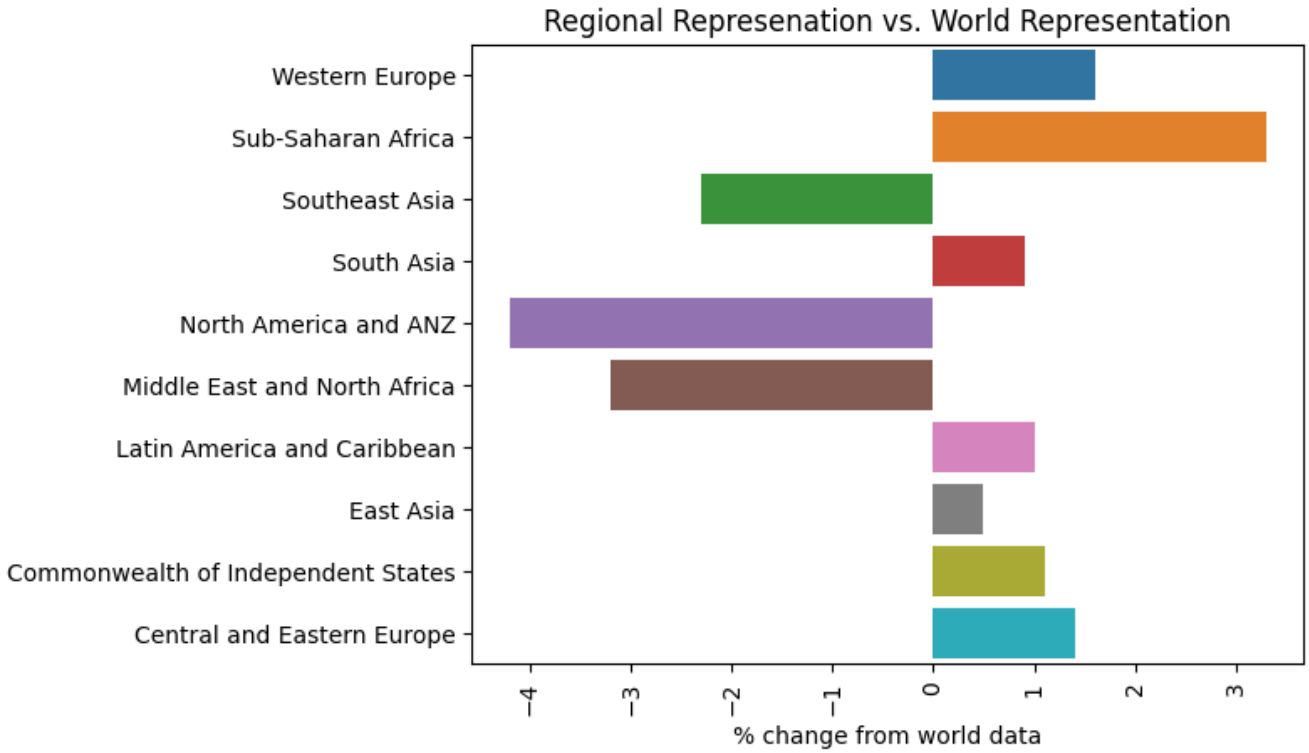
Regional Representation

For regions, we used a list of countries from Encyclopedia Britannica to obtain a list of all the countries in the world. We then compared the countries in our dataset with this comprehensive list and determined what percent of each region of the world we had represented in the data. Some manual matching of countries to regions was needed during this step. See EDA notebook for more information on how this was implemented.



The data set did not have complete data for any single region. All regions were missing some countries. This was further complicated by the fact that some regions would be expected to contain a higher percentage of the world's countries, even if the data were 100% complete.

This question was addressed by calculating the percentages we would expect each region to take up with complete data, and then comparing those percentages to the percentages in the data we had.



Based on this analysis it was determined that there was no overt regional representation issue in the dataset. The data were not perfect, but were not so far off that any additional corrective steps were deemed necessary.

2.3 Baseline Modeling

2.3.1 Random Forest Regression

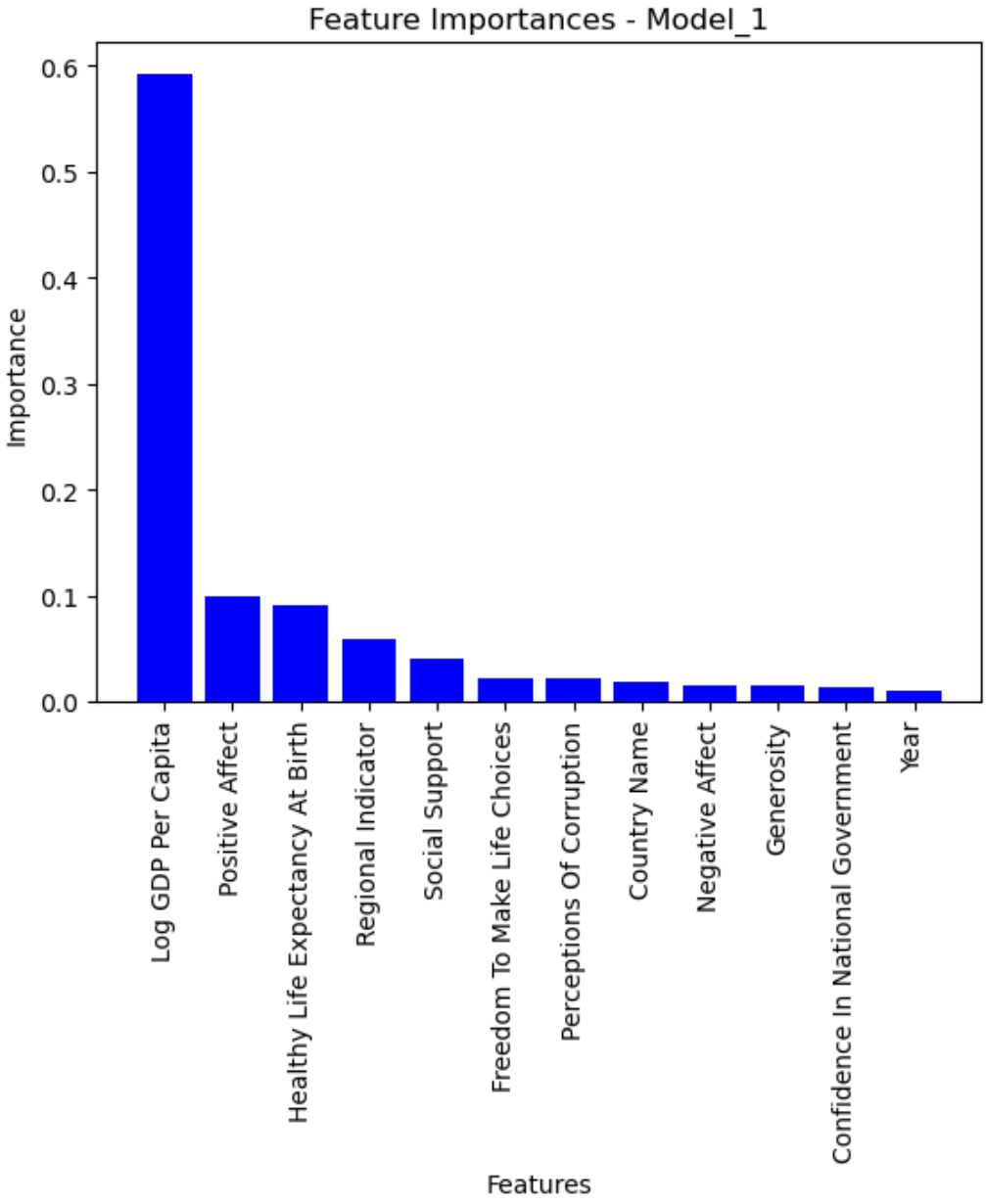
Baseline modeling and feature selection work was done with a random forest regression model with no hyper parameter tuning.

Several identical RF Regression models were trained and tested on a variety of feature combinations to gain an understanding about which feature combinations might result in the best overall prediction of the Life Ladder Target. The various subsets of features that were used for training these models were based on the five main features that had been previously shown to have the highest correlation with the target variable during EDA.

Model	**Algorithm**	**# Features Used**	**r2 score**	**MAPE**
Model 1	RF Regression	12	0.8985	0.0559
Model 2	RF Regression	9	0.8822	0.061
Model 3	RF Regression	3	0.8161	0.0736
Model 4	RF Regression	5	0.8589	0.0669

Feature importances were also plotted for all models. The feature importance plot for the best performing model (Model_1) is shown below.

After comparing model performance, it was decided that all of the data features would be used in modeling since the model trained on all of them had a slight edge in performance. Feature importances were plotted for all models. The feature importance plot for the model trained on all of the features (Model_1) is shown below.



The results of this testing indicated that three of the features could be dropped for only a slight performance reduction in the out of the box model. If the dataset had been larger, this would have been a tempting tradeoff, offering still good performance with a large reduction in the feature space and required compute time.

Given the small number of features, and the small relative size of the data set, it was determined to keep all the features in play, since the amount of compute needed to build a model like that was not an issue.

Some exploration of the n_estimators hyper-parameter for the RF Regression model was also done at the baseline modeling stage (see Preprocessing_and_Training_RF notebook).

2.3.2 KMeans Clustering

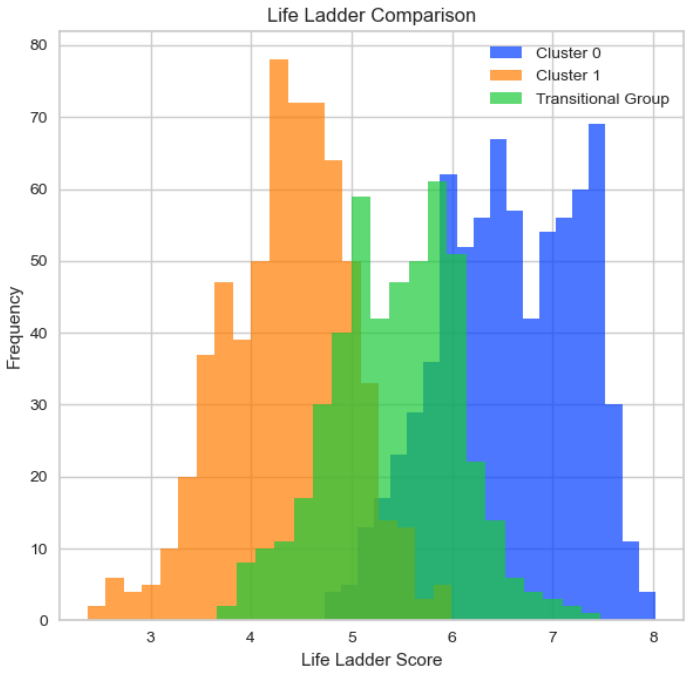
In addition to the RF Regression model. A KMeans model was made, grouping the data into two relatively clear clusters and evenly sized clusters.

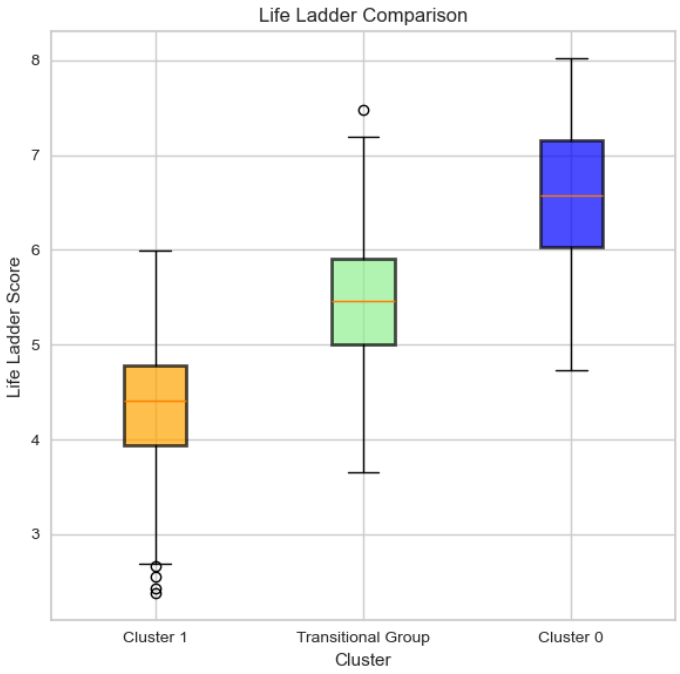
Mean values for the numerical features in the respective clusters are shown below:

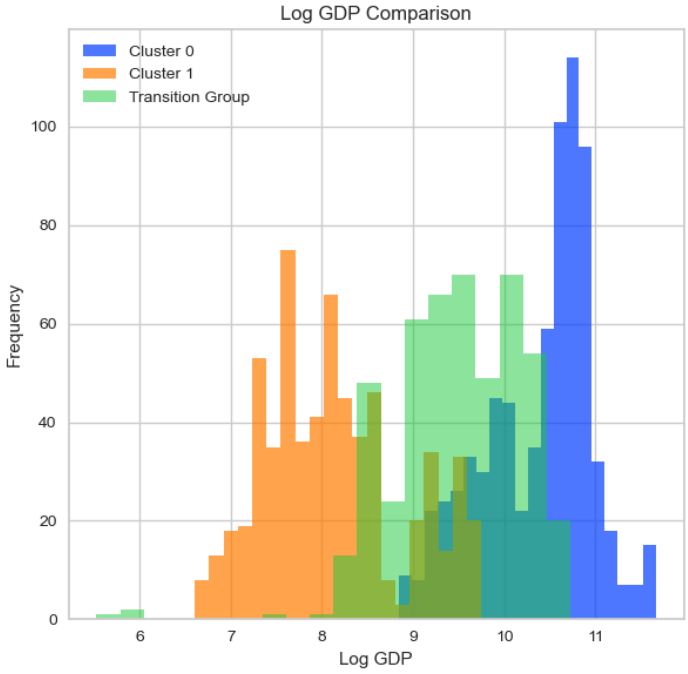
Feature	**Cluster 0**	**Cluster 1**
Life Ladder	6.377	4.543
Log GDP	10.189	8.461
Social Support	0.892	0.725
Healthy Life Expectancy	68.429	57.687
Freedom To Make Life Choices	0.823	0.666
Generosity	0.029	-0.024
Perceptions Of Corruption	0.689	0.812
Positive Affect	0.713	0.599
Negative Affect	0.245	0.296
Confidence In National Government	0.460	0.502

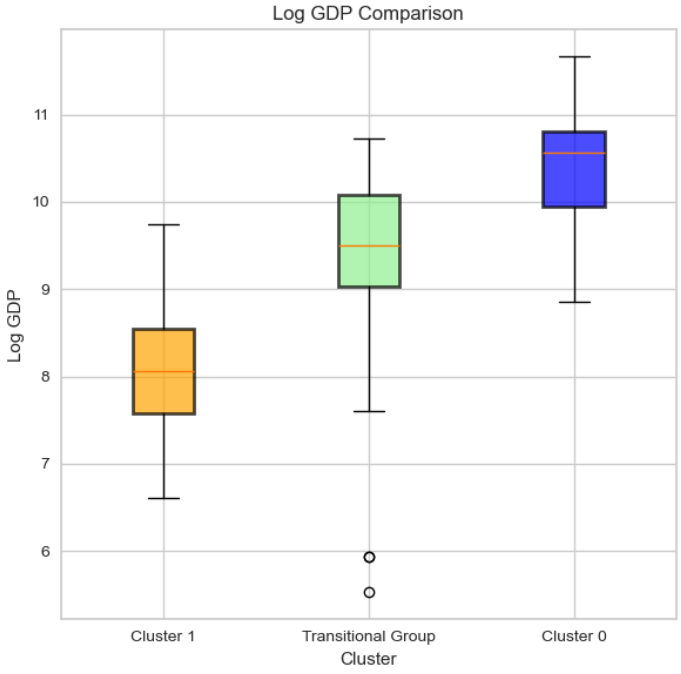
Some of the countries in the data set were also observed to switch from one cluster to another depending on the year. These countries were isolated into a third transitional group. The clusters were further analyzed in the context of the most important features as identified in the RF Regression model.

For each of these features, we noted separation between the clusters and a clear area of transition. The following page provides examples of the type of separation observed between the groups in this case for the target variable and for GDP. Similar separations were noted in each of the major features examined.

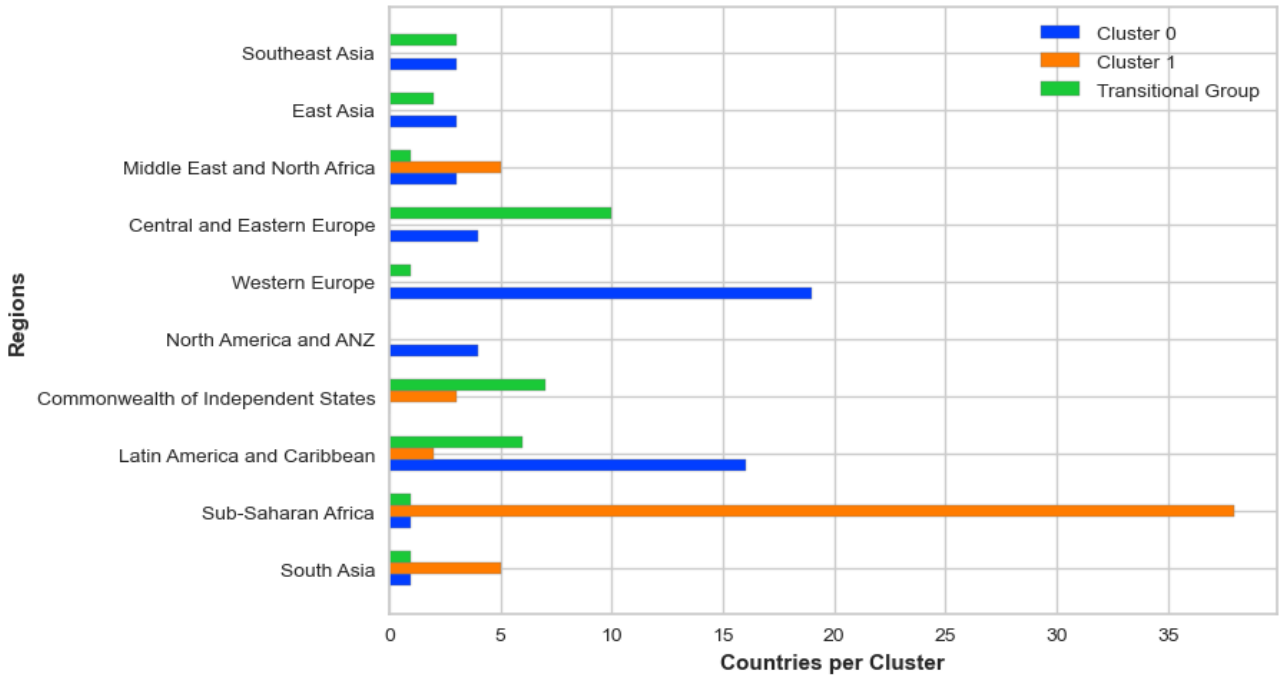

Life Ladder Comparison
Cluster 0
Cluster 1
Transitional Group
Frequency
Life Ladder Score


Life Ladder Comparison
Life Ladder Score
Cluster 1
Transitional Group
Cluster 0
Cluster


Log GDP Comparison
Cluster 0
Cluster 1
Transition Group
Frequency
Log GDP


Log GDP Comparison
Log GDP
Cluster 1
Transitional Group
Cluster 0
Cluster

Regional representations of each of the clusters are provided below to provide additional insight into how the clusters are distributed globally.



The K-means model was not used or developed for prediction. Instead, it was meant to help us gain a deeper understanding/intuition about the structure of the data. The zones of transition identified within the KMeans model will appear again below, where they provide additional insight into the core question of the project.

2.4 Extended Modeling and SHAP analysis

In the final phase of modeling, we performed hyper-parameter tuning on the RF regression model mentioned above, as well as an XGBoost regression model and a LightGBM model. These models were chosen as comparisons with the RF model because they were also tree-based but had the gradient boosted component, which it was hoped might allow us to capture additional nuance in the data.

All three models were trained on the data in their out-of-the-box configurations and then tuned in an effort to improve performance. The performance metric chosen for comparison was MAE as seen below.

Model	**r2 score**	**MAE**	**MAPE**	**MAE Improvement over untuned model**
RF Regression	0.8978	0.2828	0.056	0.78%
XGBoost	0.9103	0.2675	0.0526	7.91%
LGBM	0.9107	0.2701	0.0533	3.67%

All three models performed well, with the XGBoost model performing slightly better than the other two. This model also showed the most improvement over the baseline, with an almost 8% decrease in MAE over an untuned XGBoost model.

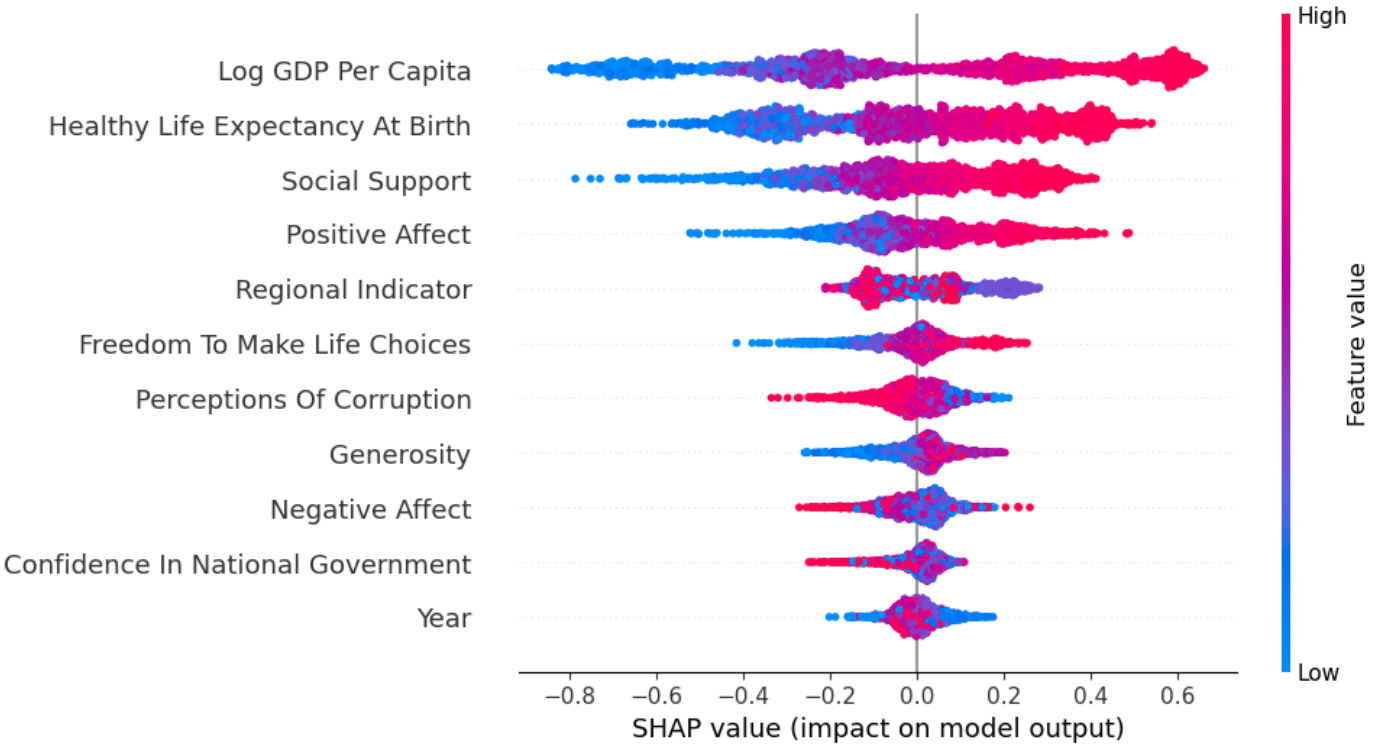
SHAP Analysis

Finally, to understand the impact of the features on the target in the trained model, a SHAP analysis was performed.

This analysis addressed the core question of the project, and allowed us to gain a deeper understanding of how the features related to the target variable, not just in terms of raw importance, but in terms of impact and direction as well.

Individual plots were also made of the five most impactful features (Log GDP Per Capita, Healthy Life Expectancy At Birth, Social Support, Positive Affect, and Regional Indicator). These plots can be found below.

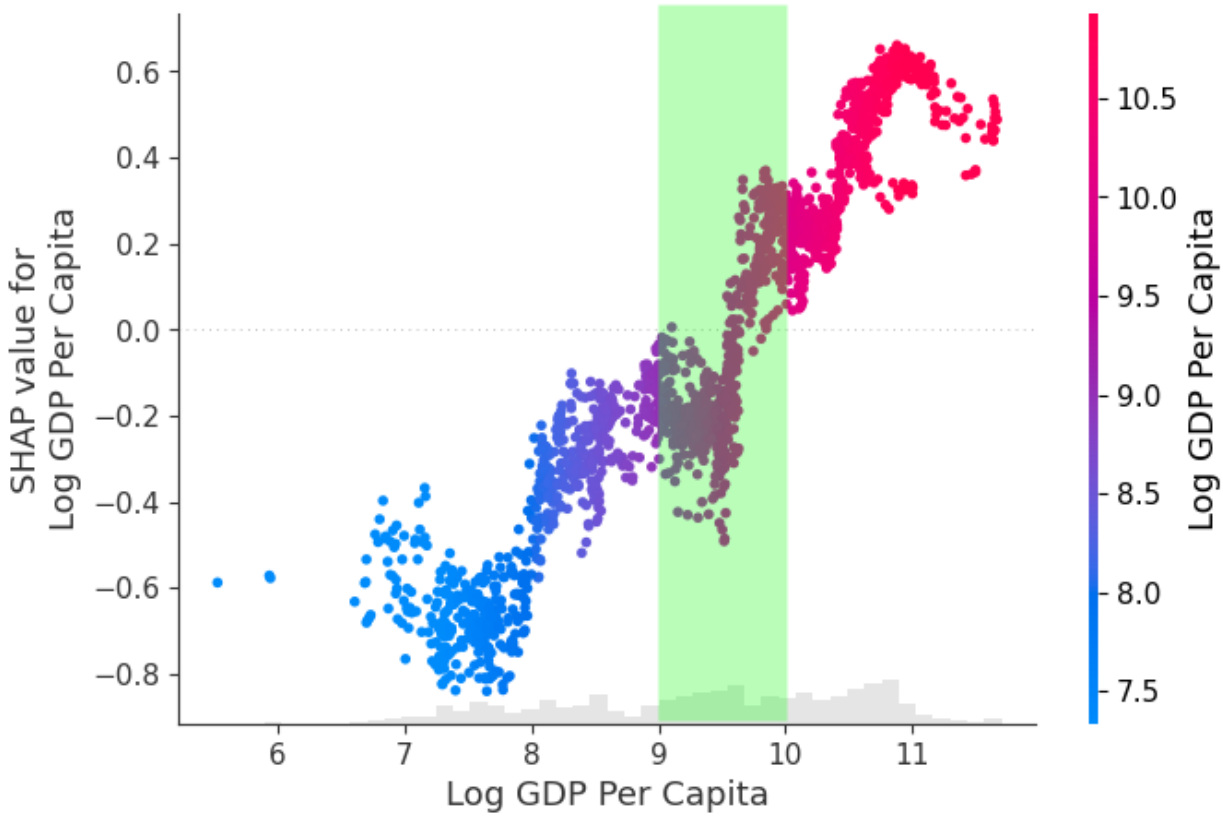
SHAP Analysis for all features

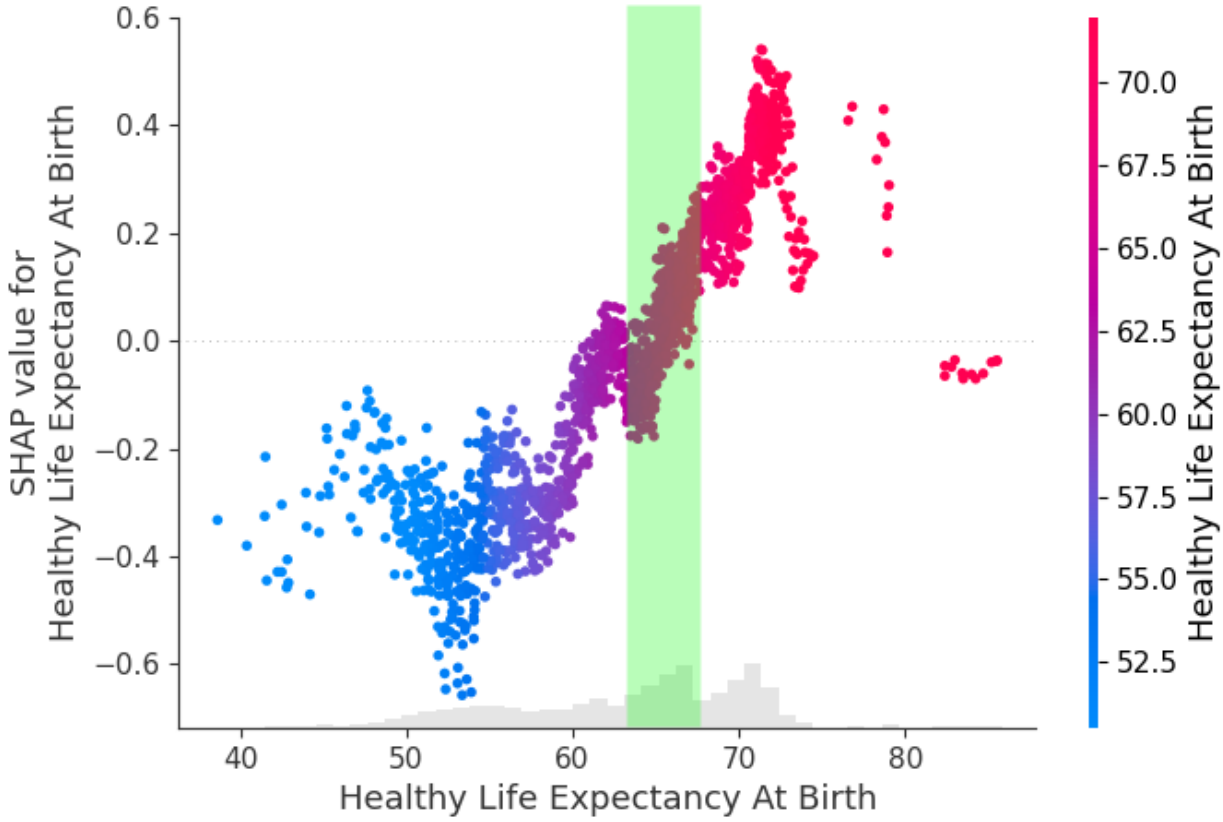


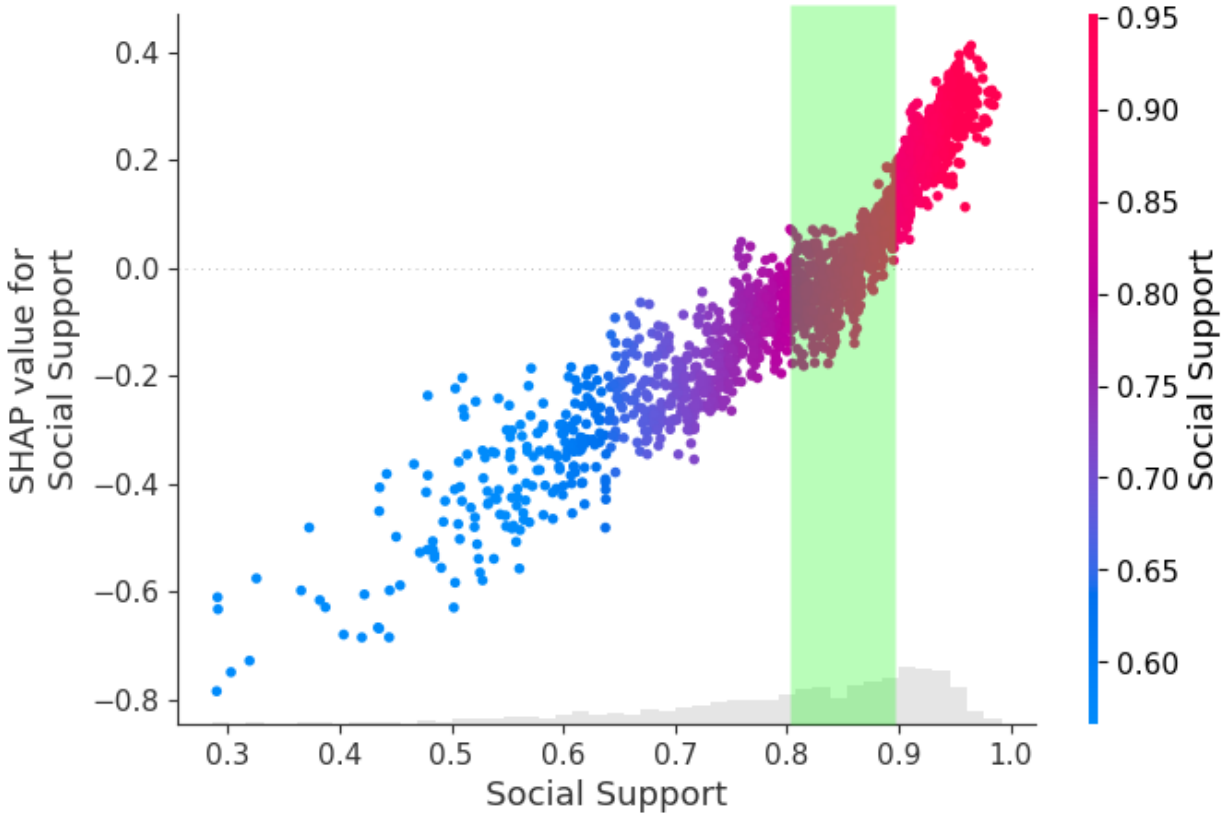
The overall ranking of the SHAP values for each feature closely matched the feature importances we observed in the various modeling steps done earlier in the project. Visualizations of the SHAP value ranges for each of the most important features are shown below in more detail.

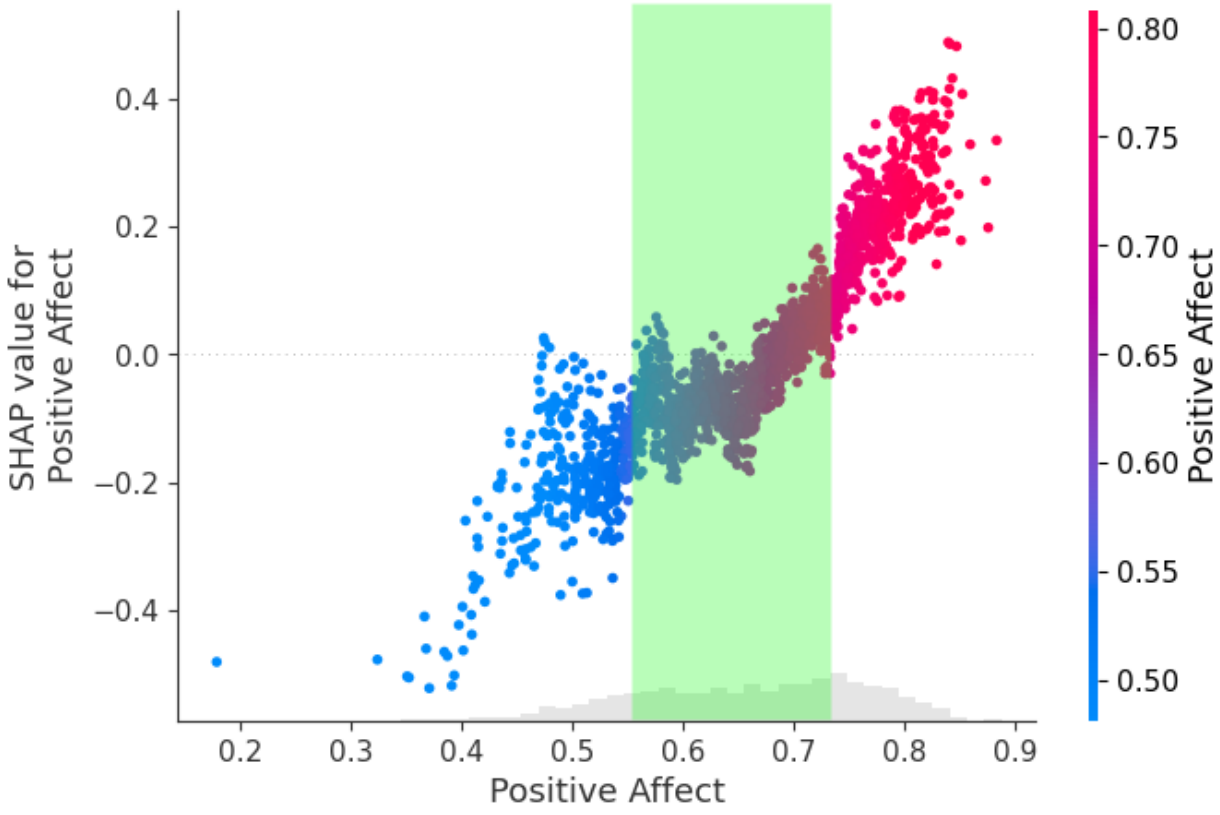
For each numerical feature modeled, there was a threshold where the impact of the feature flipped from being negative to positive. Within the individual features, these thresholds were noted to correspond very closely with the transition zones in the K-Means analysis, indicating that the countries in these transition zones were exhibiting feature values in the range where they were starting to have a positive impact on the target variable, pushing the country from cluster 1 to cluster 0.

The green overlays below indicate the approximate middle quartiles of the KMeans transition zones.

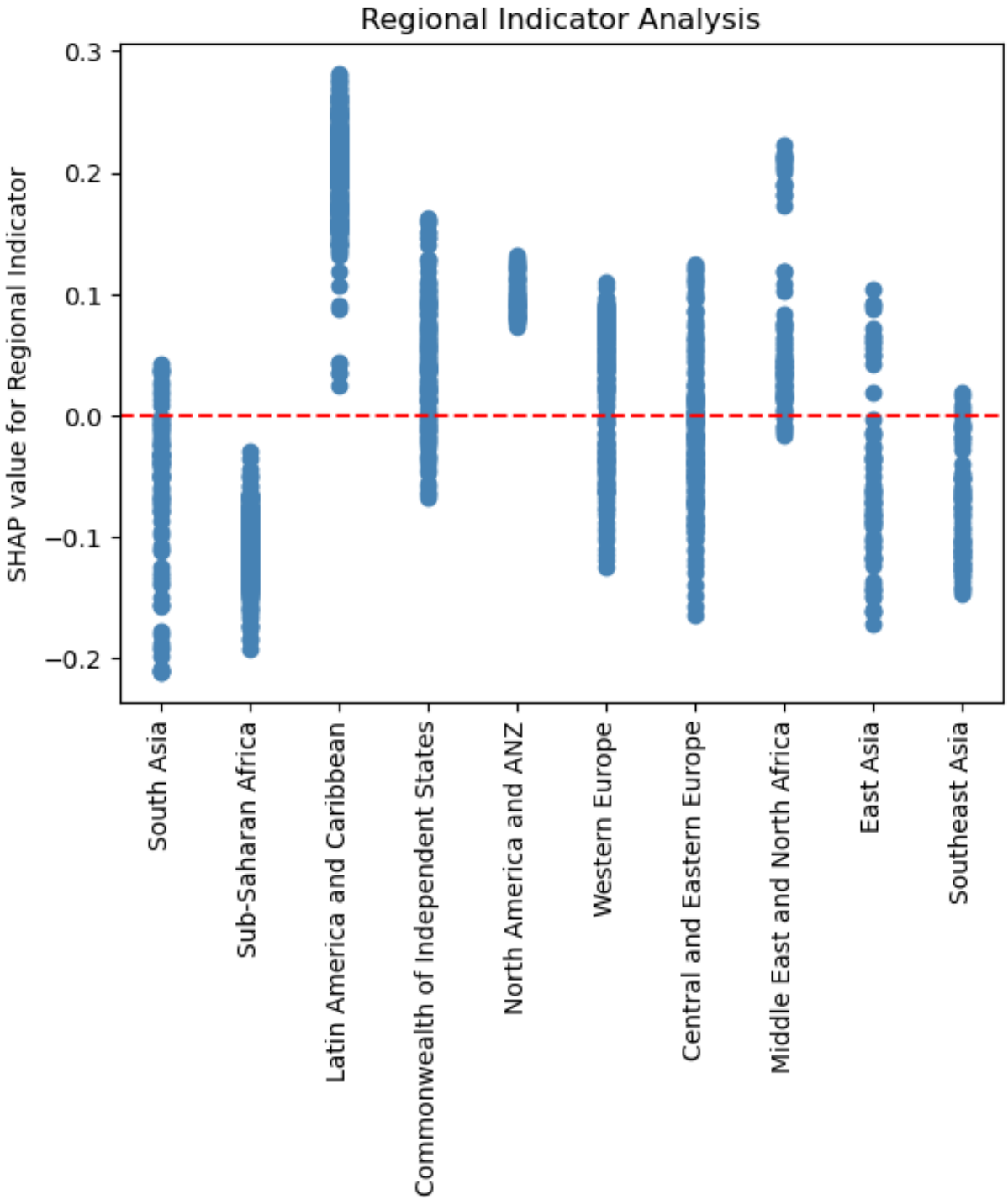

SHAP value for
Log GDP Per Capita
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
6
7
8
9
10
11
Log GDP Per Capita
10.5
10.0
9.5
9.0
8.5
8.0
7.5
Log GDP Per Capita


SHAP value for
Healthy Life Expectancy At Birth
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
40
50
60
70
80
Healthy Life Expectancy At Birth
70.0
67.5
65.0
62.5
60.0
57.5
55.0
52.5
Healthy Life Expectancy At Birth


SHAP value for
Social Support
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Social Support
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
Social Support


SHAP value for
Positive Affect
0.4
0.2
0.0
−0.2
−0.4
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Positive Affect
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Positive Affect

Region also plays a role based on the SHAP results, with both Latin America and the Middle East and North Africa having a small positive impact in the target. This impact seems to be decoupled from GDP and to a lesser extent from life expectancy, given that other regions that perform better on these metrics have less SHAP impact.



4. Conclusions and Future Work

The ranges presented in the SHAP analysis can provide a guide to stakeholders interested in understanding or affecting the perceptions of life satisfaction in a population as measured via the life ladder scores.

Based on the model developed above, pushing a few key metrics into or past their transition zones would be expected to have a positive effect on a population’s perception of how generally good their lives are. That effect would be expected to grow generally stronger the higher the scores are pushed above the thresholds.

We can't say there is causation based on this model, but we can gain an intuition for the relationships. GDP is a major component in terms of importance and impact when looking at its effect on the Life Ladder scores in our data. It's significant, but it's not the whole story. Multiple other features come into play, some of which have outsize negative or positive impacts depending on the feature value. The major features to consider are shown below along with approximate threshold values beyond which a positive impact on the target would be expected. Regional indicator is fixed and is therefore not included below.

Feature	**Measures**	**Approximate threshold**
Log GDP	Economic Strength	9.8
Healthy Life Expectancy	Life Span	67 years
Social Support	Strength of Personal Social Networks	0.9
Positive Affect	Previous days enjoyment	0.7

Next steps

Future work on the data provided by the World Happiness Survey should try to address more granular levels of analysis. This project is focused on the globe, using all the countries we have data for in this data set.

Both regional- and country-level modeling would shed additional light on dynamics that may be obscured by the global approach.

Country-level modeling was not possible in this project given the small number of data points per row in this data set, and it may not be possible until much more data from the Happiness Survey is available.

Regional modeling may be a more promising approach and an avenue for future analysis.The finding that some regions have an impact on the target variable that seems to be uncorrelated with their economics is interesting and bears further scrutiny.

5. Recommendations

The modeling process in this project suffered because some data points, in particular data points for the Perceptions of Corruption and Confidence in National Government features were not available for all countries and all years.

This seems to be because conducting national surveys of populations on these questions is not possible in some areas and some political structures. This limitation meant that some very significant countries could not be modeled because there was not complete information in the data set.

We know that all the features are important in the sense that removing any of them seems to lessen model performance to some degree. However, it is also clear from the analysis of both the feature importances and the SHAP values that some features are much more important and impactful than others.

The model would be improved if data from those key features could be used for all countries, without totally sacrificing the problematic features. For example, the missing government and corruption feature values could be imputed through a new regression model, or a proxy feature could be developed that captures the same sort of information on various governments without having to actually ask the survey questions.

Development of a proxy value, and remodeling of the data set using that value, seems the best approach since it can be based on outside data that is presumably more transparent than values produced by a regression model and used for imputation.

6. Consulted Resources

https://worldhappiness.report/

https://data.worldbank.org/indicator/SP.DYN.LE00.IN?end=2021&start=1960&view=chart

https://www.britannica.com/topic/list-of-countries-1993160