Google Capstone Project

World Happiness Report

James Lynch

November 2024

Contents

1 Overview

This project serves as the Capstone project for the Google Data Analytics Professional Certificate. It explores the World Happiness Report dataset, examining different models for predicting a country's happiness score based on the data provided in the dataset.

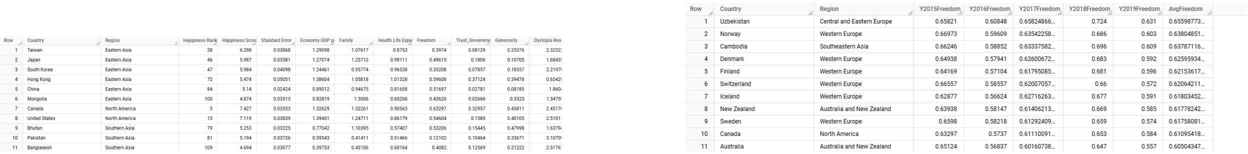
2 Objectives

This project aims to build different models—Linear Regression, Multiple Linear Regression, k-Nearest Neighbors, and Decision Tree—to accurately predict countries' happiness scores, and analyze these models to understand why certain models performed better than others.

3 Data Cleaning

3.1 SQL Cleaning

The first step of this project was to create multiple tables for each column in the tables pictured below. The resulting tables contain columns with the happiness scores for every country in a given year, and the average of each country's happiness score over the period in separate columns.



Row	Country	Region	Happiness Rank	Happiness Score	Standard Error	Economy GDP p	Family	Health Life Expe	Freedom	Trust_Governme	Generosity	Dystopia Res
1	Taiwan	Eastern Asia	38	6.298	0.03868	1.29098	1.07617	0.8753	0.3974	0.08129	0.25376	2.3232
2	Japan	Eastern Asia	46	5.987	0.03581	1.27074	1.25712	0.99111	0.49615	0.1806	0.10705	1.6843
3	South Korea	Eastern Asia	47	5.984	0.04098	1.24461	0.95774	0.96538	0.33208	0.07857	0.18557	2.2197
4	Hong Kong	Eastern Asia	72	5.474	0.05051	1.38604	1.05818	1.01328	0.59608	0.37124	0.39478	0.6542
5	China	Eastern Asia	84	5.14	0.02424	0.89012	0.94675	0.81658	0.51697	0.02781	0.08185	1.860
6	Mongolia	Eastern Asia	100	4.874	0.03313	0.82819	1.3006	0.60268	0.43626	0.02666	0.3323	1.3475
7	Canada	North America	5	7.427	0.03553	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811	2.4517
8	United States	North America	15	7.119	0.03839	1.39451	1.24711	0.86179	0.54604	0.1589	0.40105	2.5101
9	Bhutan	Southern Asia	79	5.253	0.03225	0.77042	1.10395	0.57407	0.53206	0.15445	0.47998	1.6379
10	Pakistan	Southern Asia	81	5.194	0.03726	0.59543	0.41411	0.51466	0.12102	0.10464	0.33671	3.1070
11	Bangladesh	Southern Asia	109	4.694	0.03077	0.39753	0.43106	0.60164	0.4082	0.12569	0.21222	2.5176

(a) Example of an uncleaned table

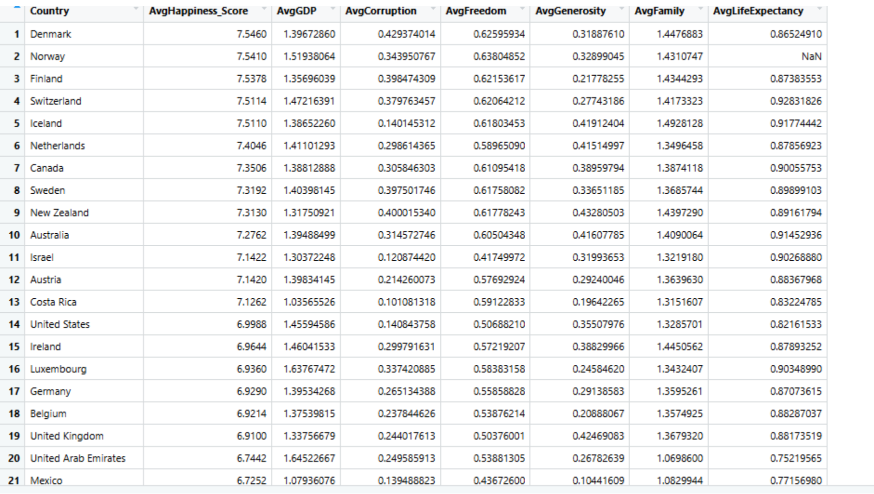
Row	Country	Region	Y2015Freedom	Y2016Freedom	Y2017Freedom	Y2018Freedom	Y2019Freedom	AvgFreedom
1	Uzbekistan	Central and Eastern Europe	0.65821	0.60848	0.65824866...	0.724	0.631	0.65598773...
2	Norway	Western Europe	0.66973	0.59609	0.63542258...	0.686	0.603	0.63804851...
3	Cambodia	Southeastern Asia	0.66246	0.58852	0.63337582...	0.696	0.609	0.63787116...
4	Denmark	Western Europe	0.64938	0.57941	0.62600672...	0.683	0.592	0.62595934...
5	Finland	Western Europe	0.64169	0.57104	0.61795085...	0.681	0.596	0.62153617...
6	Switzerland	Western Europe	0.66557	0.58557	0.62007057...	0.66	0.572	0.62064211...
7	Iceland	Western Europe	0.62877	0.56624	0.62716263...	0.677	0.591	0.61803452...
8	New Zealand	Australia and New Zealand	0.63938	0.58147	0.61406213...	0.669	0.585	0.61778242...
9	Sweden	Western Europe	0.6598	0.58218	0.61292409...	0.659	0.574	0.61758081...
10	Canada	North America	0.63297	0.5737	0.61110091...	0.653	0.584	0.61095418...
11	Australia	Australia and New Zealand	0.65124	0.56837	0.60160738...	0.647	0.557	0.60504347...

(b) Example of table for a given variable

Figure 1: Uncleaned vs Cleaned Table

3.2 R Cleaning

The next step in the data-cleaning process was to extract the average value for each variable from the tables I created. I calculated the average for each variable over the four-year period to normalize the data points for all countries, ensuring that outliers in a particular year would not skew the results. This was accomplished in R using join statements to generate a single table containing the average value of each variable for each country.



	Country	AvgHappiness_Score	AvgGDP	AvgCorruption	AvgFreedom	AvgGenerosity	AvgFamily	AvgLifeExpectancy
1	Denmark	7.5460	1.39672860	0.429374014	0.62595934	0.31887610	1.4476883	0.86524910
2	Norway	7.5410	1.51938064	0.343950767	0.63804852	0.32899045	1.4310747	NaN
3	Finland	7.5378	1.35696039	0.398474309	0.62153617	0.21778255	1.4344293	0.87383553
4	Switzerland	7.5114	1.47216391	0.379763457	0.62064212	0.27743186	1.4173323	0.92831826
5	Iceland	7.5110	1.38652260	0.140145312	0.61803453	0.41912404	1.4928128	0.91774442
6	Netherlands	7.4046	1.41101293	0.298614365	0.58965090	0.41514997	1.3496458	0.87856923
7	Canada	7.3506	1.38812888	0.305846303	0.61095418	0.38959794	1.3874118	0.90055753
8	Sweden	7.3192	1.40398145	0.397501746	0.61758082	0.33651185	1.3685744	0.89899103
9	New Zealand	7.3130	1.31750921	0.400015340	0.61778243	0.43280503	1.4397290	0.89161794
10	Australia	7.2762	1.39488499	0.314572746	0.60504348	0.41607785	1.4090064	0.91452936
11	Israel	7.1422	1.30372248	0.120874420	0.41749972	0.31993653	1.3219180	0.90268880
12	Austria	7.1420	1.39834145	0.214260073	0.57692924	0.29240046	1.3639630	0.88367968
13	Costa Rica	7.1262	1.03565526	0.101081318	0.59122833	0.19642265	1.3151607	0.83224785
14	United States	6.9988	1.45594586	0.140843758	0.50688210	0.35507976	1.3285701	0.82161533
15	Ireland	6.9644	1.46041533	0.299791631	0.57219207	0.38829966	1.4450562	0.87893252
16	Luxembourg	6.9360	1.63767472	0.337420885	0.58383158	0.24584620	1.3432407	0.90348990
17	Germany	6.9290	1.39534268	0.265134388	0.55858828	0.29138583	1.3595261	0.87073615
18	Belgium	6.9214	1.37539815	0.237844626	0.53876214	0.20888067	1.3574925	0.88287037
19	United Kingdom	6.9100	1.33756679	0.244017613	0.50376001	0.42469083	1.3679320	0.88173519
20	United Arab Emirates	6.7442	1.64522667	0.249585913	0.53881305	0.26782639	1.0698600	0.75219565
21	Mexico	6.7252	1.07936076	0.139488823	0.43672600	0.10441609	1.0829944	0.77156980

Figure 2: The Final Table with an average column for each variable for each country

4 Taking a Deeper Look

Now that the data has been cleaned and merged into one table, we can finally start to look into what makes a country have a high happiness score.

4.1 Top Correlation

To gain a better understanding of what affects a country's happiness score, I first looked at how much each variable is correlated to the happiness score. From this, we can see that GDP, life expectancy, and family are better predictors of happiness scores while freedom, corruption, and generosity are not as good predictors.

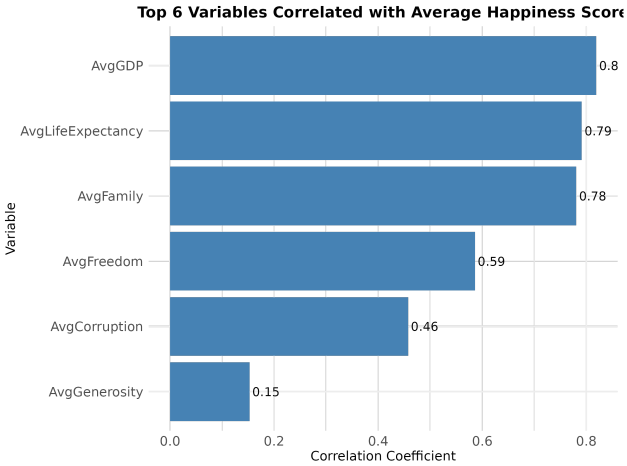


Figure 3: Top variables correlated with the happiness score

This next chart further illustrates the correlation of a certain variable to a country's happiness score by drawing a line of best fit through the given variables scatter plot. As we can see the variables with the highest correlation have lines that nicely fit the data while the variables with a lower correlation have lines that don't fit the data.

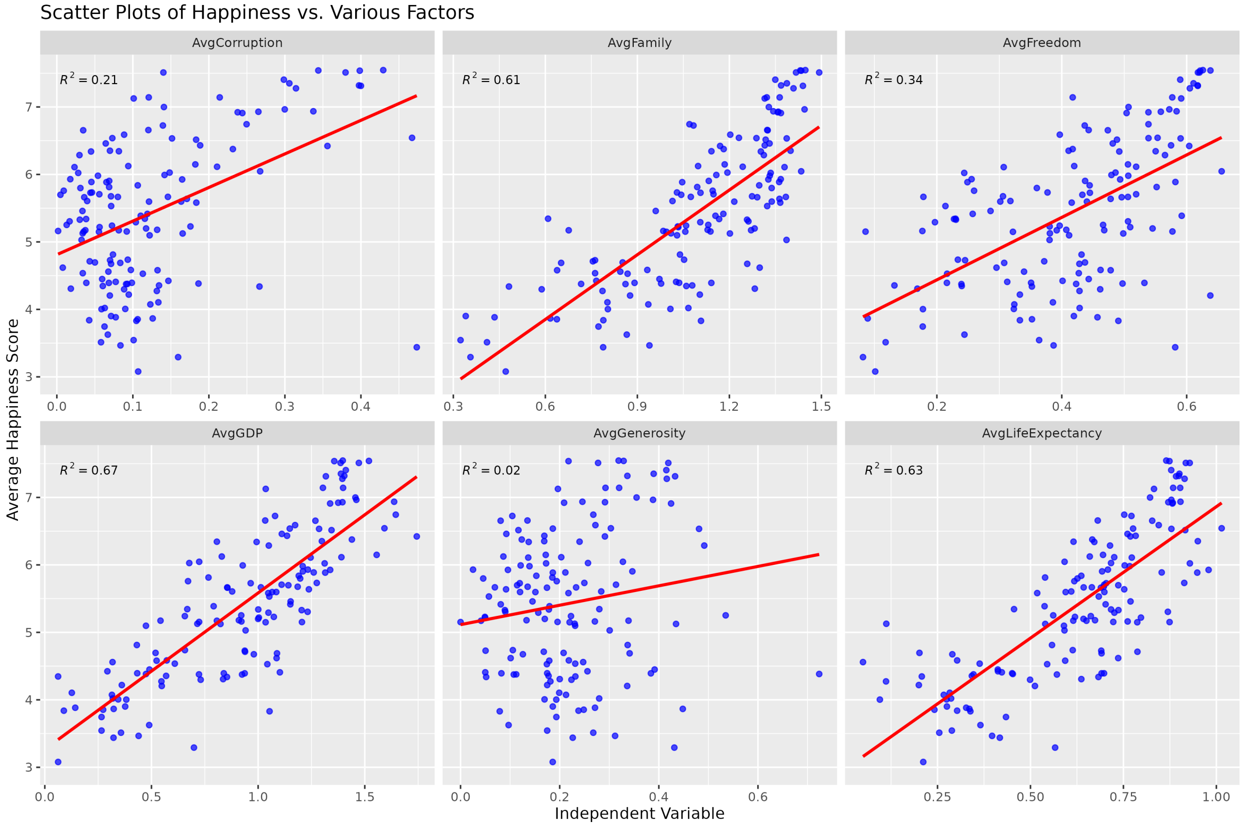
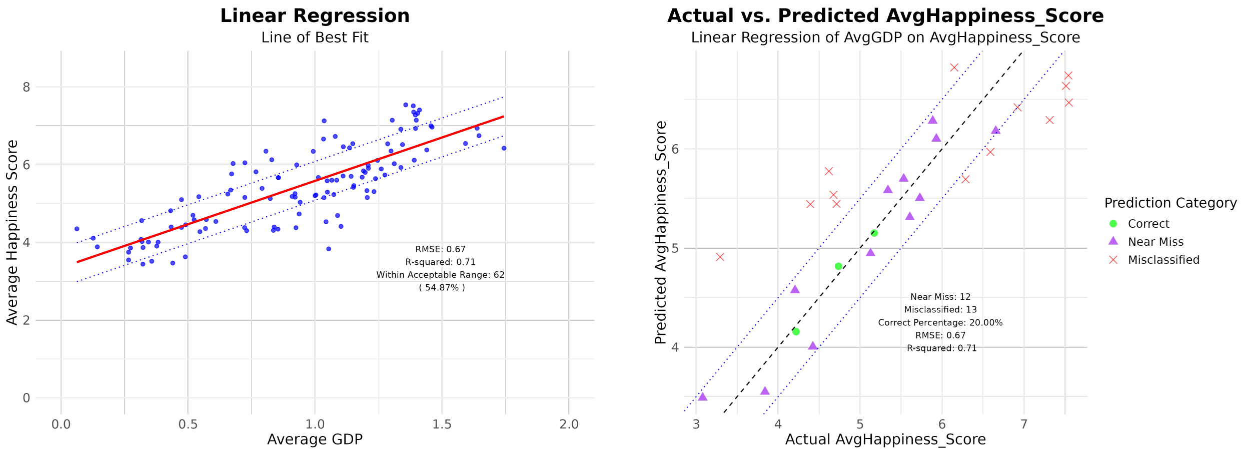


Figure 4: Line of best fit for each variable

5 Linear Regression

5.1 Approach

From the previous graphs, it is clear that GDP has the strongest correlation with happiness score. Because of this, the first way I choose to try to model the happiness score is with a simple linear regression to see how well I could predict the happiness score with one variable.



(a) Line of Best Fit

(b) Actual vs Predicted

Figure 5: Linear Regression Model Results

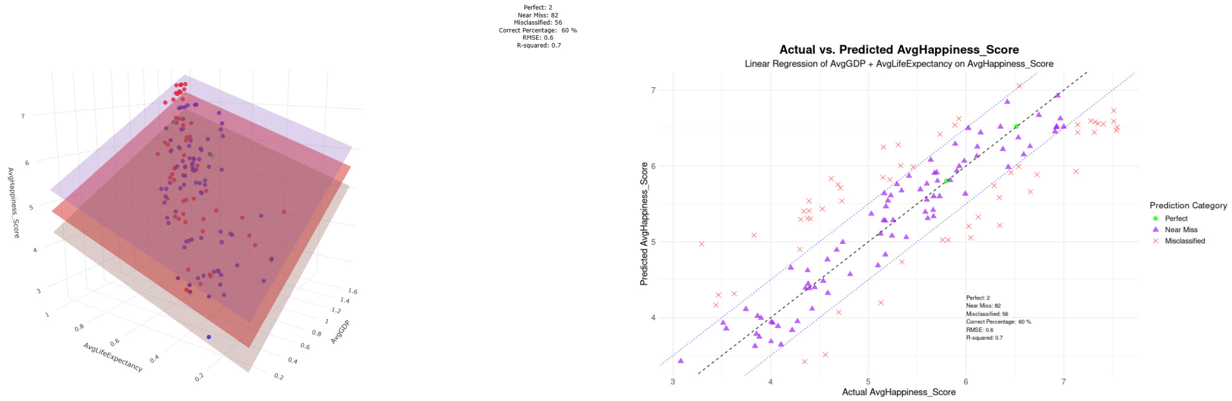
5.2 Results

Looking at the scores for this model we can see that while it might be able to get within a certain range of accuracy, it fails to consistently predict the happiness score of the country within a reasonable ±0.5, and has a not significant R^2 value of 0.71 and relatively high RMSE of 0.61. These results are indicative that happiness of a country is determined by more than one variable.

6 3D Regression

6.1 Approach

Since attempting to model with only one variable failed, we will attempt to model the data with a multiple linear regression model using two variables, while trying to see if using an interaction term helps the model. To choose the models to build I chose GDP and life expectancy as they are the highest correlated.



(a) 3D Model Plane

(b) Actual vs Predicted

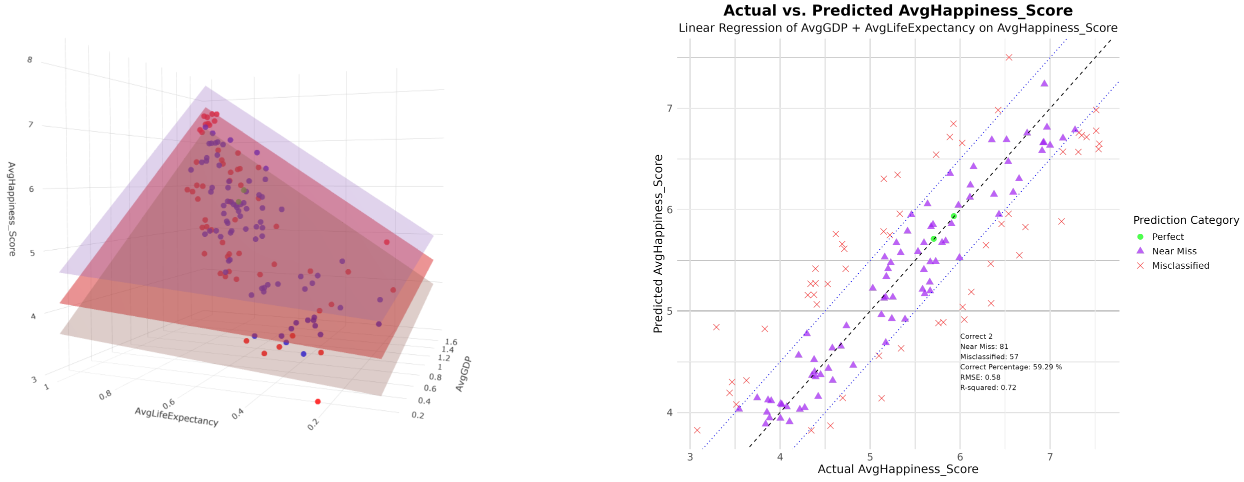
Figure 6: 3D Regression Model Results

6.2 Results

This model is a step up from the last model at correctly predicting a country within the range of error, and the model's RMSE of 0.6 is a slight improvement from 0.67. However, its R^2 value is actually 0.1 less than the linear regression model.

6.3 Accounting for Interaction

The lack of improvement from the model and worse R^2 could be attributed to the lack of accounting for the interaction between the two independent variables. To account for this we use interaction terms which account for how much the change of one independent variable affects the change of another independent variable, and thus how much that change affects the output.



(a) 3D Model with Interaction

(b) Actual vs Predicted

Figure 7: 3D Regression Model with Interaction Results

6.4 Results

Analyzing the model with the interaction term, we can see that it is marginally better than a model without the interaction term, as it has an ever so slightly lower RMSE and a higher R^2 term. These statistics indicate that the plane drawn is a slightly better fit than the linear model, but can we do better?

7 Multiple Linear Regression

7.1 Approach

For multiple linear regression, instead of just looking at one or two of the independent variables, this time we look at all of them at once and build two separate models where one accounts for interaction and one doesn't.

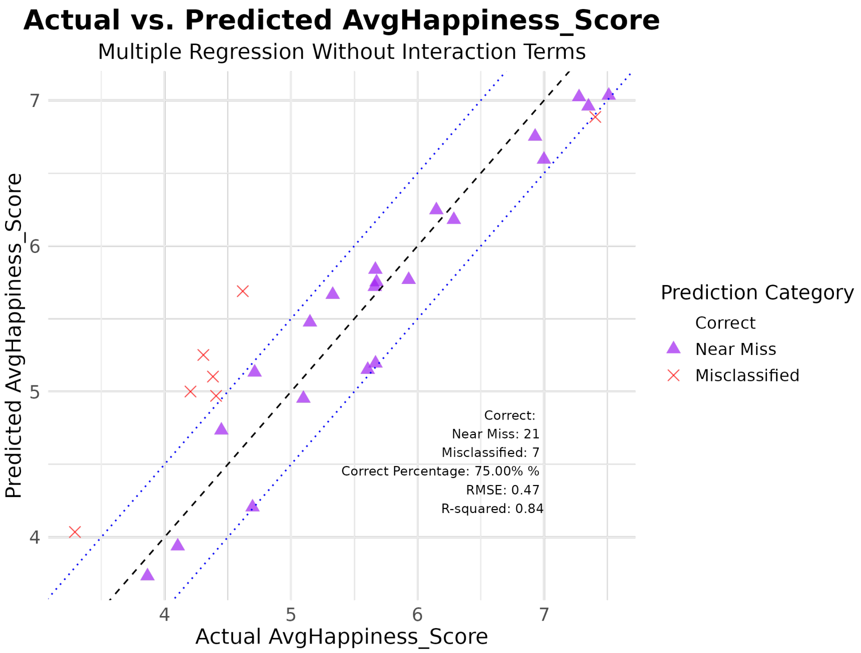


Figure 8: Actual vs Predicted Values Without Interaction

7.2 Results for Model Without Interaction

This model shows significant improvement in our abilities to guess the happiness score of a given country by accounting for all variables. The RMSE and R^2 show significant improvement, as the R^2 of 0.84 approaches the value of statistical significance, and the RMSE score means that each guess is just barely on average within our arbitrary error bounds.

7.3 Interaction

When adding interaction to this model, I decided to only use the pair-wise interaction terms. A pair-wise interaction term is only how much one variable correlates to exactly one other variable. Non-pair-wise terms include three or more terms. I chose only to include pair-wise terms as adding too many interaction terms causes the models to overfit the data.

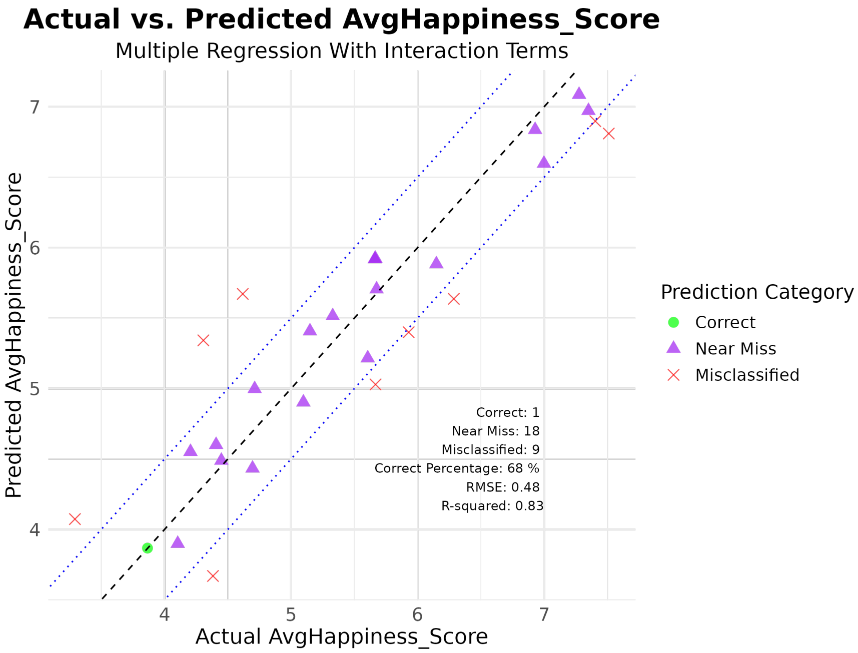


Figure 9: Actual vs Predicted Values With Interaction

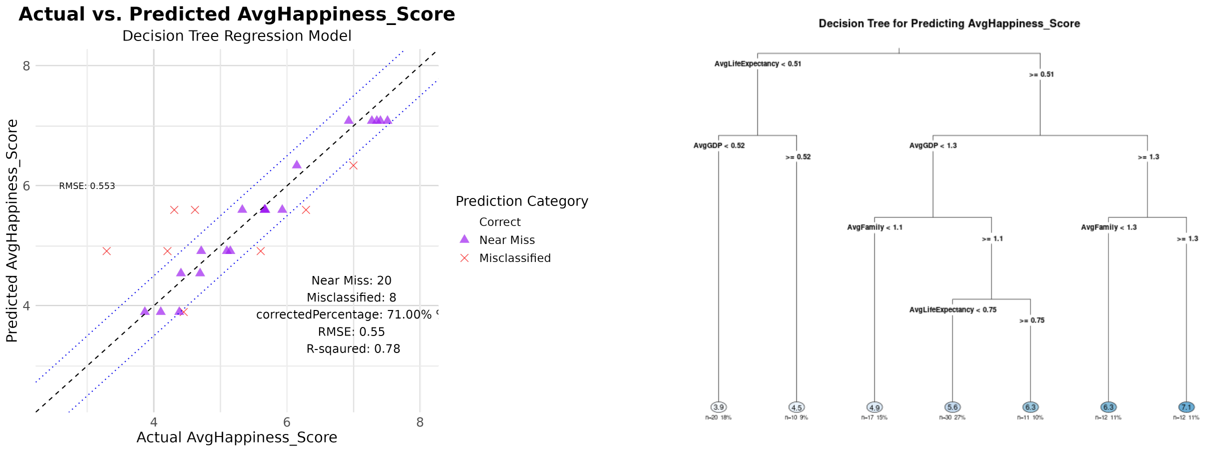
7.4 Results for Interaction Model

The results for the interaction model show practically no improvement, as it has the same RMSE and a slightly lower R^2 value. This lack of improvement is a result of the type of model we are using for the given dataset. Given that we are calculating such a complex value, directly trying to predict the happiness score by trying to build a function might not be the best approach. Instead, trying to use clustering models or more abstract ways of regression might achieve better results.

8 Decision Tree

8.1 How it Works

A decision tree regression model predicts a value by constructing a tree-like structure where each internal node represents a decision based on the value of a specific feature, where at its final layer it classifies it into a group and guesses a score based on the scores of the other data points in its group. To build the tree you have to have training data where the model finds the decision point where the sum of the residuals is the lowest.



(a) Actual vs Predicted

(b) Decision Tree Model

Figure 10: Decision Tree Model Results

8.2 Results

The results from this, while not an improvement from the multiple linear regression model without interaction, still did well at predicting the happiness score with an RMSE of 0.55 and an R^2 of 0.78.

9 k-Nearest Neighbor

9.1 Approach

The k-nearest neighbor model builds off the last models by not directly trying to find a continuous function for predicting a value but by trying to group data points. The k-nearest neighbors achieve this by finding the k nearest happiness scores based on the countries with the most similar independent variables to the unknown country. Then it finds the average of those happiness scores.

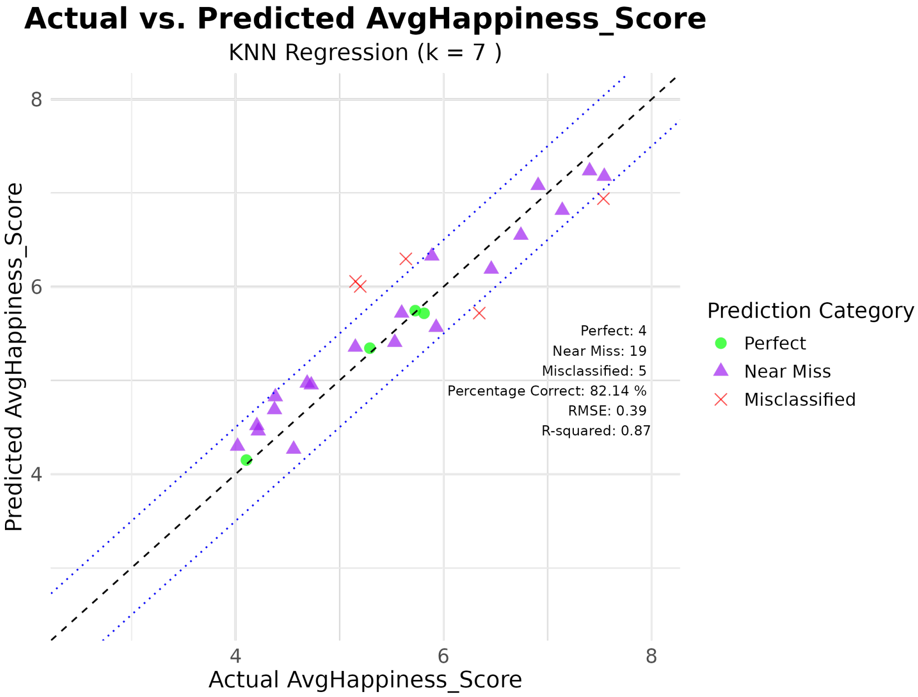


Figure 11: Actual vs Predicted

9.2 Results

This model shows a significant ability to guess a country's happiness score and is by far the best-performing model. An RMSE of 0.39 and an R^2 of 0.87 show that this model guesses within a tight range of the actual value. This model performs so well because of its methodology of using a group of most similar countries to guess the happiness score instead of looking at all the countries. This does better for this task as even with six independent variables, the number of factors that play into the well-being of a country's people is very complex.

9.3 Further Proof of Explanation

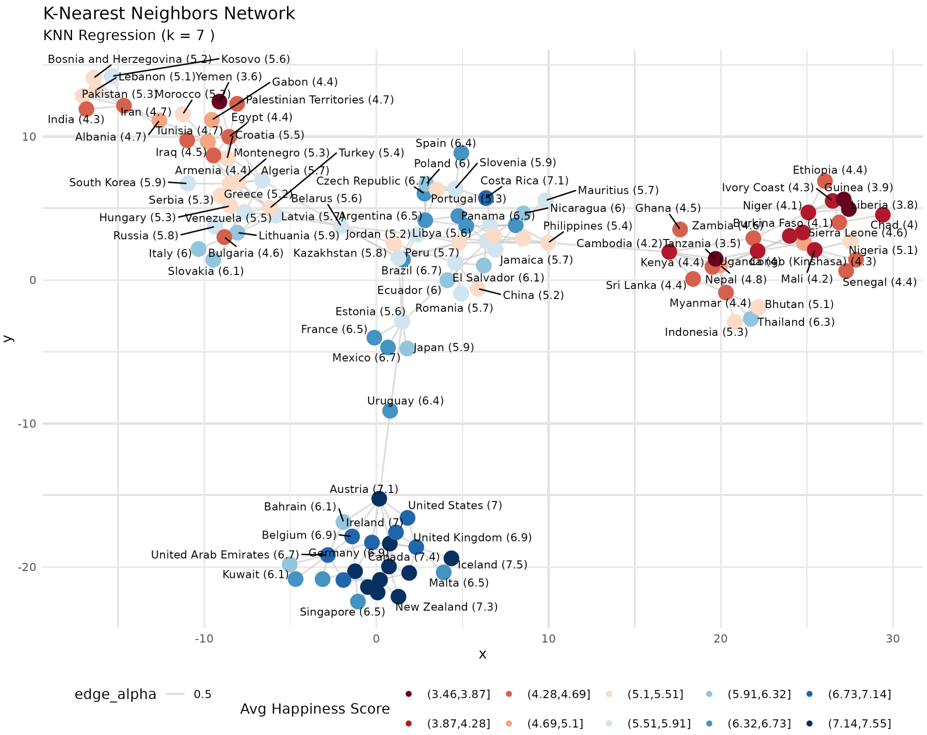


Figure 12: KNN Network Graph

9.4 Explanation of Graph

This graph shows the groupings the model made when guessing the happiness score for the test data. Each connection shows if a country was chosen as similar to another. Nodes with seven connections show the test countries while any node with fewer were associated with a test country. The coloring of the country reflects the happiness score, with the bluer the country the happier, and the redder the country the more unhappy it is.

9.5 Analysis of Groups

Looking at the groupings that the model made, we can see that the groups it generates reflect real-world conditions. The group of the bluest countries consists of wealthy Western countries, rich oil states, and tax havens. When looking at the upper left grouping, we can see that group consists of countries that are or were in a state of armed conflict, and when looking at the rightmost group we can see that it has associated a bunch of African countries with each other. Finally, when looking at the middle group we can see countries in the world middle class or from more emergent markets in the world.