**Global Happiness Report**

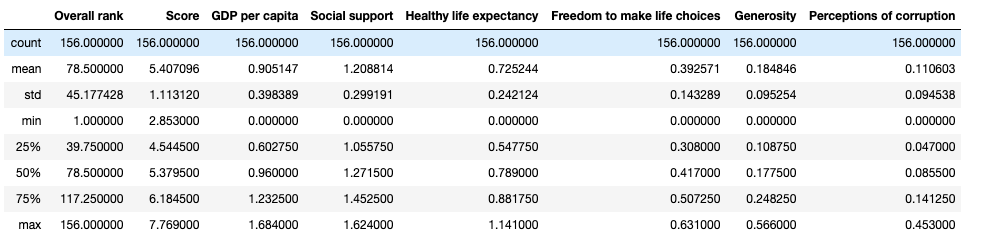
**Ian Johnson, Hayden Dessommes, Ethan Kaiserman**

**Our Dataset:**

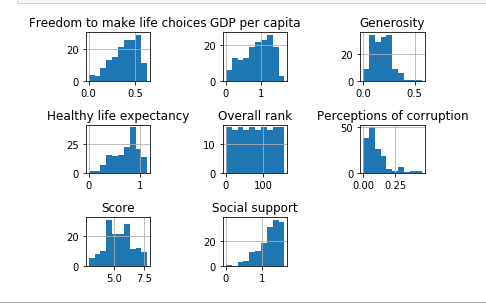
We chose to use a dataset that contains different predictors to determine a countries overall happiness rating for their average citizen

**Features/Variables:**

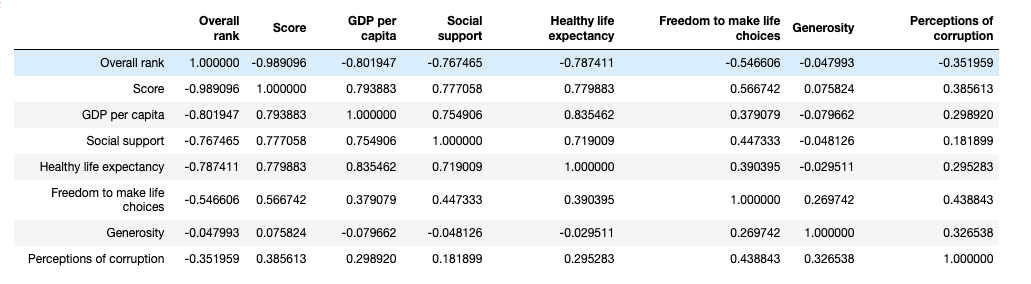
* GDP per capita
* Social support
* Healthy life expectancy
* Freedom to make life choices
* Generosity
* Perceptions of corruption

**Summary Statistics**

**Distribution of Features:**

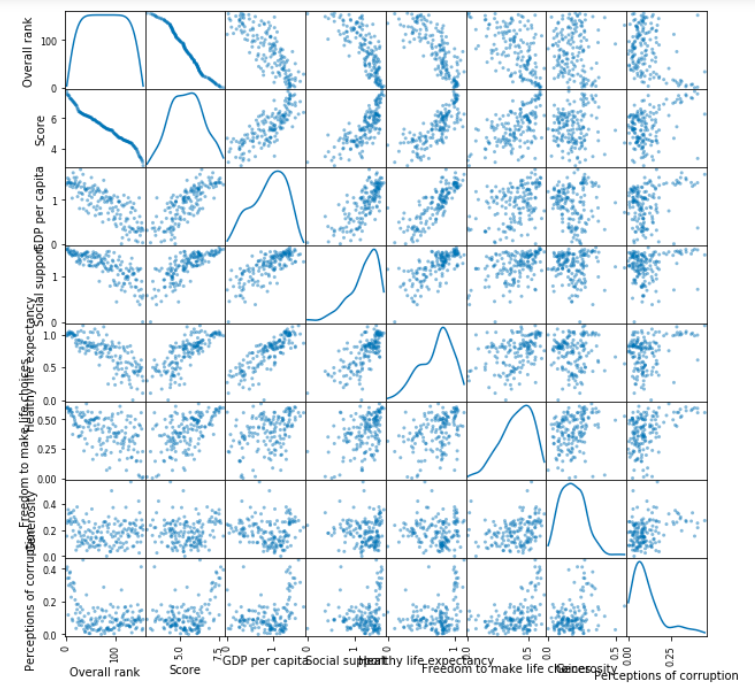


**Correlation Matrix:**

****

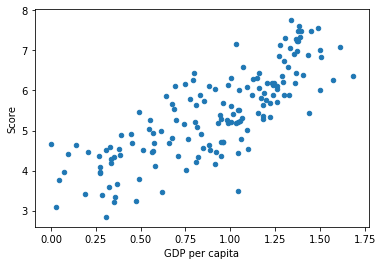
By looking at the correlation matrix, we discovered that GDP per capita, Social support, and Healthy life expectancy had the highest correlation with overall score. To further investigate this we created scatter plots comparing score with each of these three variables.

**Correlation Scatter Plot:**

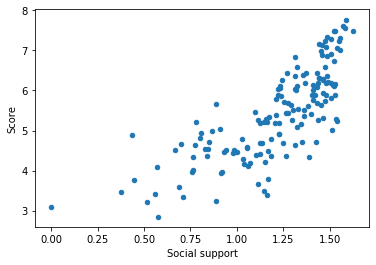
****

**Scatterplots:**

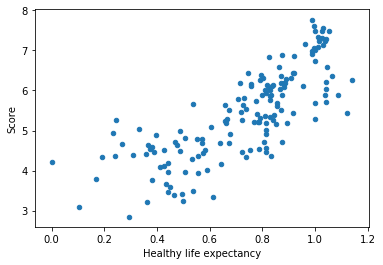
**GDP per Capita vs Score:**



**Social Support vs Score:**



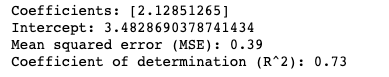
**Healthy life expectancy vs Score :**

****

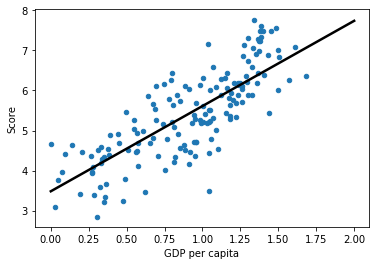
**Linear Regression Model:**

* Split the data with 80% going to the training set and 20% going to the testing set
* Started with just GDP per capita since it had the highest correlation with score

**Regression Model with just GDP per capita:**



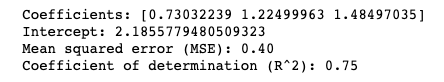
**Scatter Plot comparing GDP vs Score with added regression line:**



Adding the other two highly correlated features: to try and get a higher r-squared value:

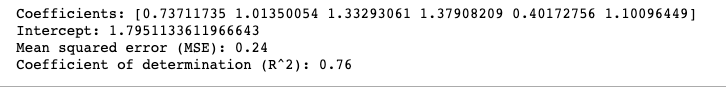
**Multiple Linear Regression Model with GDP per Capita, Social Support, and Healthy life expectancy:**

Obtained the following results:



**Multiple Linear Regression Model with all Features:**

Obtained the following results:



This model performed the best of all linear regression models using all of the variables.

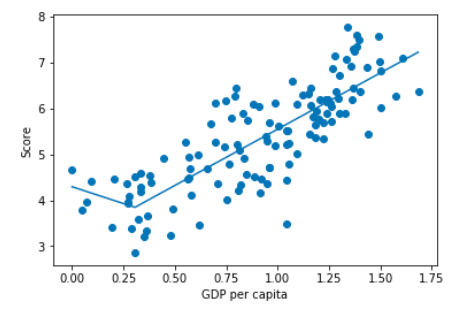
**MARS (Multivariate Adaptive Regression Splines)**

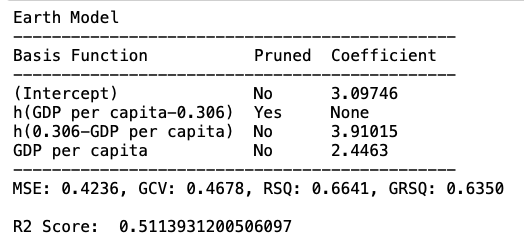
Mars is a lot like multiple linear regression, but a hinge can be inserted in the middle of the model

Here is a test on a single variable MARS model on the training data to illustrate this hinge of the model:

**MARS with One Variable**

Here we use GDP per capita, since it had the highest correlation with score out of the features.



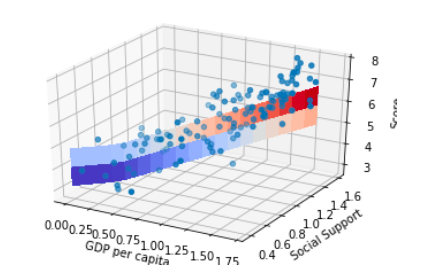
****

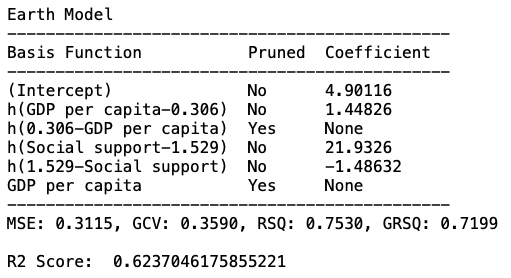
You can see the output of the summary for the MARS model gives the intercept, the hinge functions, and the x value where the graph hinges (in this case it is .306).

We can see a two dimensional model does rather poorly on this data for the R2 score. Let’s add another feature.

**MARS with Two Variables**

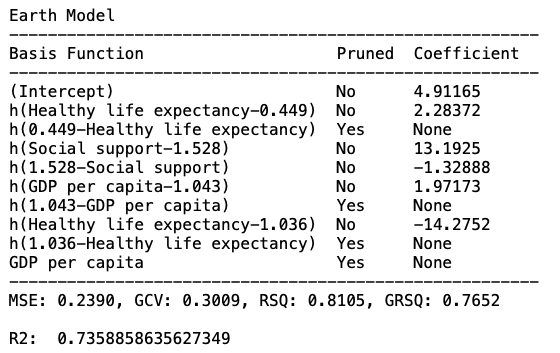
Here we use GDP per capita and Social support as our two features to test for Score





The R2 Score is better here, but we can do better.

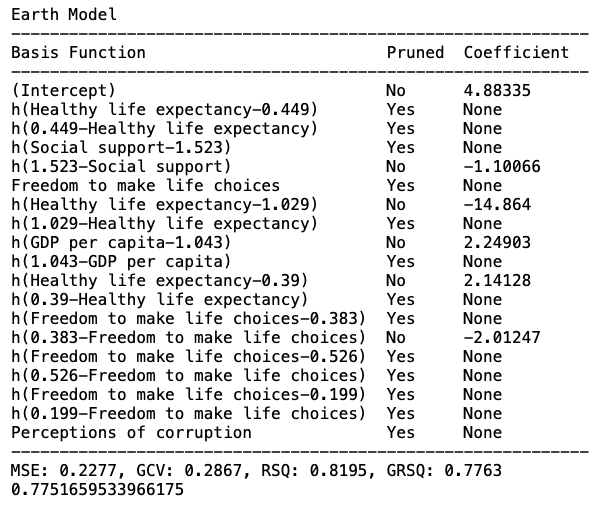
**MARS with Three Variables**

****

In this example we use the three features with the highest correlations: GDP per capita, Social support, and Healthy life expectancy.

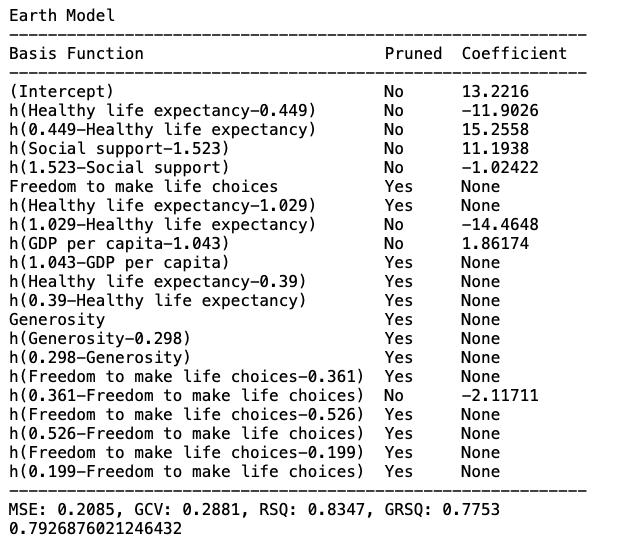
The R2 score is better, however we want to maximize the potential of MARS

**MARS with All variables But Generosity**

****

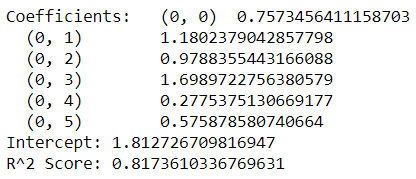
We want to see if the model might outperform using all variables if we drop the feature which has the least correlation (Generosity). It does the best so far, but let’s see if it does outperform using all variables.

**MARS with All Variables**

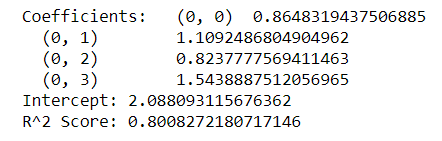
****

MARS does the best with all of the variables for our dataset. This is somewhat surprising since generosity has very little correlation with the happiness score, however there is some correlation so this makes sense.

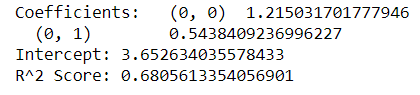
**Lasso Model:**



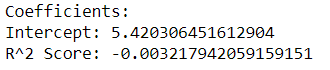
First alpha was .001. This gave us the highest R^2 value for this model, as well as the project as a whole. In this model, all coefficients are included in the model, because they are all good fits at this level.



The second alpha used was .01, which removed the last two coefficients, generosity and perceptions of corruption.



The next alpha used was 0.1, and only two coefficients remain in the model. The R^2 takes a sharp decline compared to the previous alpha. We can reasonably assume that the first few variables are weighed more in predicting Happiness compared to the last few, since the higher numbered coefficients are the first to be removed.



At alpha=0.5, no coefficients remain, and the R^2 score does not get any higher as alpha goes up. Any alpha 0.5 and higher will also have zero coefficients.

Because at the alpha=0.001 level has the highest R^2 score of this project, we can reasonably conclude that the lasso model is the best model used to fit this data set.