Megan Holford

Can You Predict Happiness?

Technical Report

Introduction:

Every year, the UN releases the World Happiness Report. This is a ranking of 155 countries in the world and how their citizens consider how “happy” they are. For the report, there are different variables that are assessed, GDP, life expectancy, generosity, social support, freedom, and government corruption. These factors are estimated to how they attribute to the overall happiness score. The data is gathered by a survey, the Gallup World Poll, of the country's population asking them to rank their happiness level by imagining a ladder with 10 steps and having the respondent answer which step or level they feel that they are on using that scale of 0 to 10. The individuals in the survey answer questions regarding their own personal situation and their happiness at that current time using the measure of that 10-step ladder.

This happiness ranking is used and accessed all over the world and many see it as a way to understand the direction of an individual countries decision making when it comes to policy and well-being for the citizens. Now of course, the data is going to be a generalized look as there will always be some that are more or less happy based on their particular situation, but looking at the country overall, it can give leaders a glimpse on what factors are most impactful to their citizens if they wanted to take that information into account when considering policy or progression.

Some working in psychology, economics, and foreign affairs even use the data collected through this report to get an indication of the progress of countries as it is done yearly since 2012, so it gives data year over year that can be used to note changes in a happiness rating and bearing in mind any policy or world changes within the last year for that country. Being able to predict a happiness score based on the different factors that it is made up from can be beneficial in many ways whether you are looking at creating or expanding new government policy, if you work in economics and want to see how GDP may relate to not only the happiness of citizens but their

health and wellbeing, or if you are in psychology and want to see how these factors relate to the idea of happiness.

Data and Background:

Using this data from the last several years, I am looking to build a prediction model to see which variables can be used to make a prediction model for the overall happiness score, and also to see which variables are more closely correlated to find if additional data can be used for this prediction.

First, I am using data from the last several years, since 2015 and since this is a yearly report, I first need to clean and combine it all to take a look at what data we have so that more data can be used for model training and testing. To clean the data, I first noticed that the column names from year to year were different for some even though it was the same data and also some years had additional columns that were not needed. I dropped the columns that I did not need and then renamed the columns for each year to be the same thing so we can combine the data. After everything was cleaned, I created a master set of the data for all years. Now I will be doing an initial analysis of the data. Upon creating some plots and visualizations, I was able to see some of the obvious correlations and relationships right away.

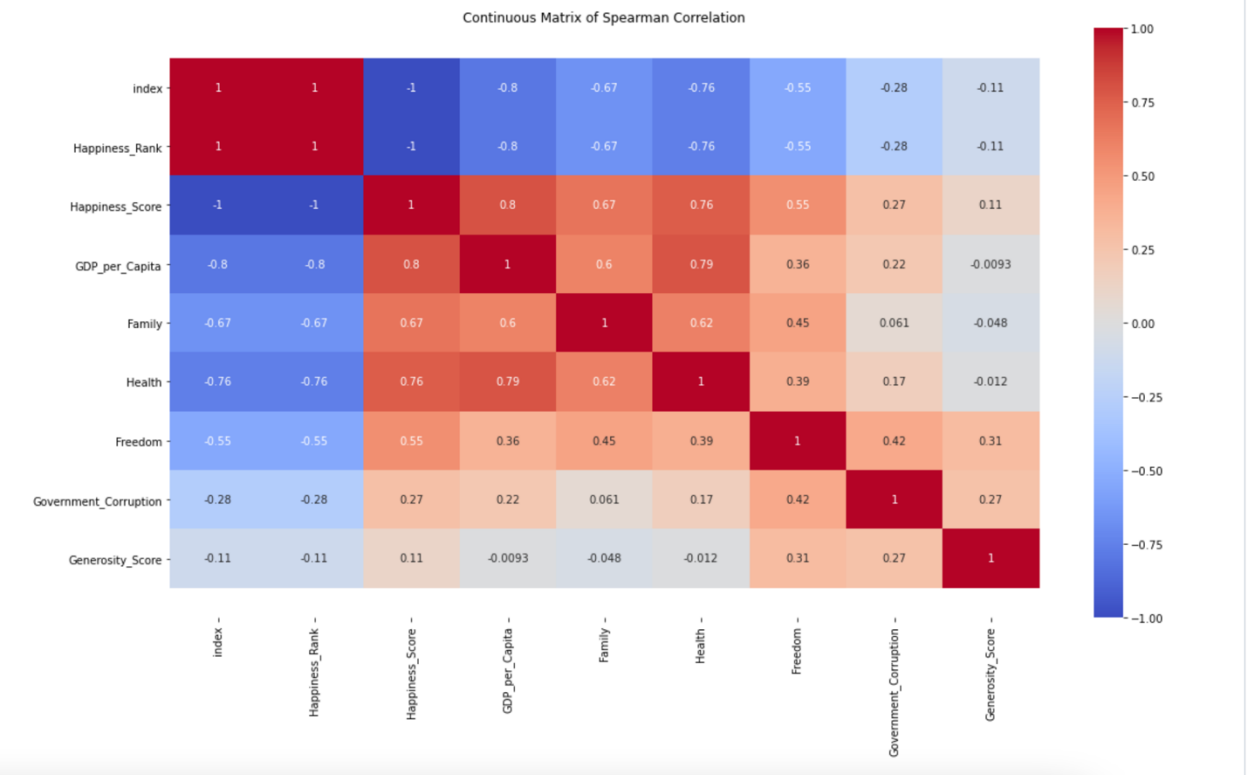

Image: Spearman Correlation heat map, showing the correlations of the variables within the data set.

This heatmap is helpful because we can see all the values and which ones have the strongest correlations and which do not have much of an impact on the overall happiness score. From this graph, we can see that the GDP and Health scores have the strongest correlation with the overall happiness score and Generosity has the weakest correlation of them all. We can also see that in the data, the ranking variable is a bit redundant for our analysis so far as it just matches the overall score, so I did remove this later.

Initial Analysis:

After noticing the correlations in the heatmap, I looked at how the values related to the happiness score individually. By using scatterplots, I could see what kind of relationship the values had together.

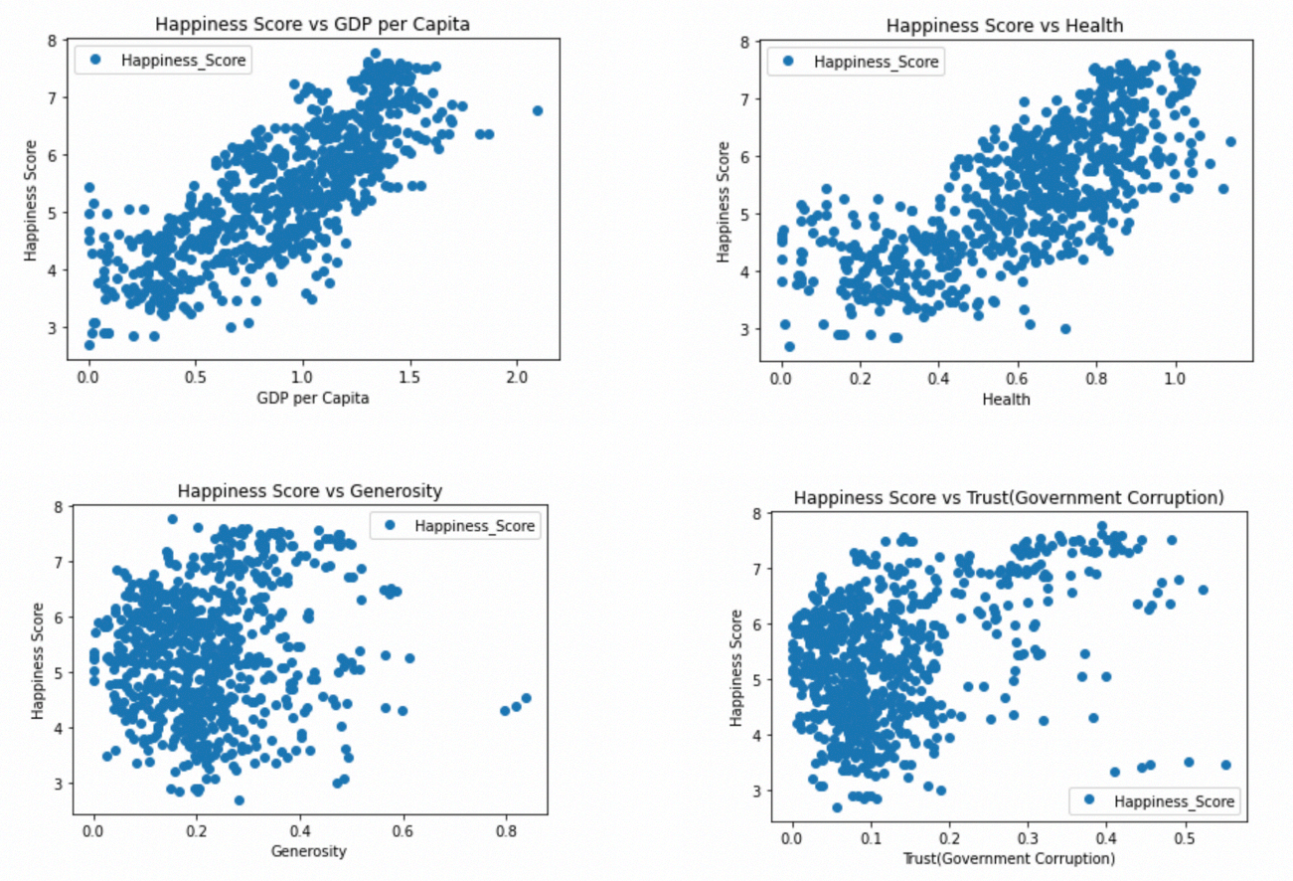

Image: Scatterplots of Happiness Score vs various values from within the data.

These charts show how the variables relate to the overall happiness and both GDP and Health have a linear relationship while I could see right away that Generosity really had no correlation.

After looking at how to different variable related to the overall happiness score, I then looked at how the scores were distributed overall. To do so, I decided to bin the full dataset using numpy linspace. I wanted to creat a high, medium, and low bin so I set it to have 4 cut offs for the bin and set the minimum value and the maximum value as the minimum and maximum scores in the whole data set. After that, I had an array that I could use to set my bins.

```
array([2.56690001, 4.3141667 , 6.06143339, 7.80870008])
```

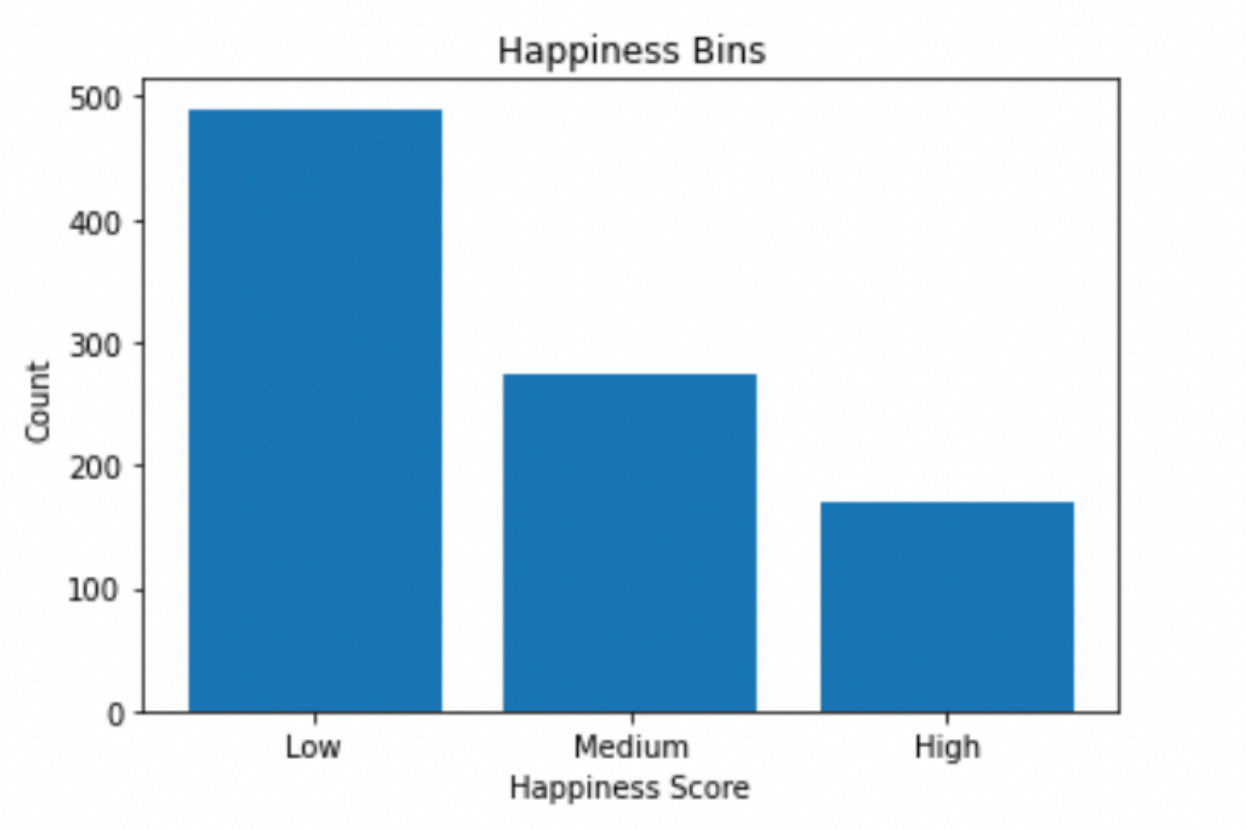

Image: Bar chart showing binned values of happiness scores in three categories (high, medium, and low)

The bar graph for the binned values showed that the happiness scores definitely had more scores in the low category present in the data. This was helpful to understand how the scores are distributed.

Model Selection and Testing:

After doing the initial analysis of the data, I began to work on the model selection. Since there was a linear relationship with some of the variables and the overall happiness score, I started with a linear regression model.

```
Training R2: 0.752105
Test R2:
0.7607134172909658
```

The training and testing scores were very close, but I wanted to try something else to see if I could get a higher r-squared score and a better fitting model.

Next, I tried a Random Forest model. I started with a small forest of 10 trees for my first run.

```
Train Dataset
: 0.9633882105707343
```

```
Test Dataset
: 0.7989974149135673
```

The score for the testing set using Random Forest did give a better score by almost 4% so I ran a larger forest of 10,000 trees.

```
Training Set
: 0.9748209369859469
```

```
Test Set
: 0.8220265668407188
```

By using more decision trees, I was able to get a final test score of 82%. I did try a larger forest of 100,000 trees, but the score was also 82%, and with the training and test scores having a difference, I was already worried about overfitting, so I kept the 10,000 tree forest to run the predictions.

Results:

After fitting the model to the dataset, it was time to run the predictions and see how they matched up.

	Actual Happiness Score	RF Predicted Score
321	7.3160	7.188211
70	5.4770	5.317565
209	5.9560	4.735064
549	5.3580	5.840924
712	5.2470	5.538325
96	4.8980	4.380109
467	3.3490	4.347665
86	5.1230	4.892685
531	5.7520	5.544323
327	7.0060	7.030900
527	5.8350	6.129259

Image: Chart of predictions vs the actual happiness score using the final Random Forest model.

Here we can see a sample of the comparisons of the actual score versus the predicted score. Most of the scores that the model predicted were within 0.4 of the actual score, but in a few cases the score could be off by up to a whole point. The majority of the predictions did fall within 0.5 of the actual score.

There does not seem to be a bias of the model overestimating or underestimating the predicted score upon first look, but I would be interested in looking further to see if there is a relationship for why a score was predicted to be higher or lower than the actual score.

Conclusion:

While my final model made predictions that were close to the actual happiness score, there is still definitely room for improvement. This is a yearly report that is done and so as more data is made available, it can be added and possibly used to help train and better the model.

As mentioned during the initial analysis, there are some factors that seem to be more closely related to the overall happiness score, so I wonder if a model using only those factors could yield better results. Looking forward I would be interested to see how those would be on their own at predicting the happiness score.

Acknowledgements and References:

Gallup, I. (2021, June 17). *Global research*. Gallup.com. Retrieved October 1, 2021, from https://www.gallup.com/analytics/318875/global-research.aspx.

World Happiness Report. (n.d.). Retrieved October 1, 2021, from https://worldhappiness.report/.

Data: https://www.kaggle.com/mathurinache/world-happiness-report#__sid=js0

Jupyter Notebook for analysis and modeling can be found here: https://github.com/mholford91/DSC-630/blob/main/Project/Final%20Milestone.ipynb