Predicting Countrywide Happiness

Ian Stonecypher

DSCI 403 FA22

12/06/22

Abstract

The aim of this project was to create a model that can accurately predict average levels of subjective well-being (or simply “happiness”) for an entire country. Data regarding happiness levels of many countries, recorded in many years, as well as several other features of those countries, was obtained from the World Happiness Report website. Next, a linear regression was performed to predict happiness levels, using seven of the other data features as predictors. The model was then improved slightly by the transformation of features to have a more linear relationship with happiness, and by using insights obtained from a k-means algorithm run on one feature. Finally, reflection was done regarding which predicting features of a country seem to have the greatest and smallest impacts on countrywide happiness levels. Considerations for ethics and future work are provided at the end.

Overview

The dataset that I have chosen for this project comes from the 2022 World Happiness Report. The data depicts the results of surveys taken in many countries around the world, and repeated for many years. The primary feature of this data is an average Cantril Ladder score from each survey, which comes from subjects being asked to rate their own life satisfaction on a score zero to ten. Following this, there are nine other variables measuring things believed by the World Happiness Report to be related to life satisfaction. A few observations are missing values for a few of these variables, but the majority of the data seems to be complete.

This dataset intrigued me because it offered the chance to understand subjective well-being as a function of various factors in a country. This was largely inspired by another class that I am taking this semester: *HASS 498: History of the 'Good Life': Aristotle to the Anthropocene*. In this class, we have explored beliefs throughout history about what makes a person happy and how people should strive to live. This has made me increasingly curious about what conditions contribute to subjective well-being around the world, and how powerful those contributions are.

My goal with this dataset was to build a very strong predictive model for the average subjective well-being within a country, as measured by average Cantril Ladder score, based on seven predictors from the dataset. I would like to know how accurately is it possible to predict a country's well-being from these seven factors alone. I would also like to determine which of these predictors is the strongest, and if any predictors could be removed from the model without it losing much of its predictive power.

Related Work

The work I hope to do on this data is similar to the work done by the World Happiness Report. In their work, they ranked all of the countries they had data on based on their average Cantril Ladder score, and attempted to explain the differences between countries using only six of the predictor variables from the dataset described above. While they did create a predictive model, the scope of their research went far beyond this model, and they placed a heavy emphasis on understanding the ways that well-being in different countries was changing over time with response to current events. My work, on the other hand, focuses solely on making the best possible model with the data. I also attempted to use the Confidence in National Government variable in my model, which the World Happiness Report did not use in theirs.

Some similar work was also done by the World Values Survey, which studies the social, religious, and other values of countries around the world and how they have changed over time. While their findings do have some implications for human well being, particularly their finding that increased perceptions of free choice have led to higher levels of happiness over time, happiness was far from their sole focus. However, because Freedom to Make Life Choices is a predictor that I used in my model, I would be curious to compare my model to theirs and see if the importance of free choice that I found is similar to the importance they found. In the future, I

might also like to try creating a predictive well-being model based on the values that they studied.

My eye was also caught by a paper by Betsey Stevenson and Justin Wolfers, titled "Economic Growth and Subjective Well-being: Reassessing the Easterlin Paradox". This paper studies the relationship between and country's subjective well-being and its GDP per capita, and finds a clear and strong positive relationship between the two. My model also factored in GDP, but my hope was to create a model that is more complex and accurate than theirs, which was shy I used many more predictors on top of this.

Data Acquisition

My dataset, from the 2022 World Happiness Report, contains results from more than 150 countries and from years ranging from 2005 to 2021. The dataset has 2089 observations total, each row representing survey findings from a certain country in a certain year. Not every country was surveyed every year, and some rows are missing a few values. There is a download link for this data at the end of this document, and it is also available to download online from the World Happiness Report website, under their 2022 report and titled "Data for Table 2.1".

The first two columns of the dataset specify the country and year of each observation, respectively. The following ten columns represent the findings for ten response variables. The 2022 World Happiness Report describes the ten response variables as follows:

- Life Ladder (hereafter abbreviated LL): This variable reports the average Cantril Ladder score in a country for a certain year. For this, subjects in a country were asked this question: "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"
- Log GDP Per-Capita (hereafter LGDP): This variable reports the natural log of a country's GDP per-capita. GDP is based on constant 2017 international dollars, and some of the most recent measurements were actually extrapolations from GDP growth forecasts. The natural log of GDP per-capita was used because the World Happiness Report found that it fit the data better than untransformed GDP per-capita.
- Social Support (hereafter SS): This variable reports the average response to the question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?" The answer *no* was represented as 0, and *yes* was represented as 1, so this variable reports averages between zero and one.
- Healthy Life Expectancy at Birth (hereafter HLE): This variable reports life expectancy, measured in years, extrapolated from data from the World Health Organization.
- Freedom to Make Life Choices (hereafter FMC): This variable reports the average response to the question "Are you satisfied or dissatisfied with your freedom to choose what

you do with your life?", with 0 meaning *dissatisfied* and 1 meaning *satisfied*. Values are therefore averages between zero and one.

- Generosity (hereafter GEN): This variable is based on the question "Have you donated money to a charity in the past month?" Average responses between zero and one were gathered. Then, in order to correct for differences in wealth, the average responses were regressed on LGDP, and the residuals of this regression are reported in the data.
- Perceptions of Corruption (hereafter PC): This variable reports the average response to two questions: "Is corruption widespread throughout the government in this country or not?" and "Is corruption widespread within businesses in this country or not?" In cases where only one question was asked, the variable reports the average response to that question. Averages are between zero and one.
- Positive Affect: This variable reports the average response to three questions: "Did you smile or laugh a lot yesterday?", "Did you experience enjoyment a lot of the day yesterday?", and "Did you learn or do something interesting yesterday?". I did not use this variable in my model for this project.
- Negative Affect: This variable reports the average response to three questions: "Did you experience worry a lot of the day yesterday?", "Did you experience sadness a lot of the day yesterday?", and "Did you experience anger a lot of the day yesterday?" I did not use this variable in my model for this project.
- Confidence in National Government (hereafter CNG): The World Happiness Report does not do a sufficient job explaining how this data was collected or what it represents precisely.

My goal with these variables was to extract the relationship between LL, measuring countrywide happiness, and some subset of the nine other variables. One limitation I had in finding this relationship was the fact that the raw data is dotted with missing entries. Many rows of the data are missing a response for one of the ten variables, so these observations could not be used, but 1747 out of the 2089 total observations were still usable. It was also necessary to keep in mind my uncertainty about what the Confidence in National Government variable represents. It would be dangerous for me to draw solid conclusions without understanding the data inside and out, so I needed to remember that during my analysis.

Preprocessing

After loading my data into ipython, I first removed the Country and Year columns from the data. This allowed me to treat the 2089 rows of data as separate observations, without having to care about which observations came from the same countries and how those countries changed over time, because that was outside the scope of this project.

Then, I looked at the correlation values (r-squared) between every pair of the ten variables. These values are shown in Figure 1 on the next page.

	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perceptions of corruption	Positive affect	Negative affect	Confidence in national government
Life Ladder	1.000000	6.117705e-01	0.509626	0.509504	0.282297	3.518184e-02	0.184338	0.260656	0.101069	0.006540
Log GDP per capita	0.611770	1.000000e+00	0.464714	0.656724	0.130712	6.990402e-07	0.118208	0.057773	0.051020	0.036215
Social support	0.509626	4.647138e-01	1.000000	0.357101	0.167379	4.866256e-03	0.049538	0.181959	0.174948	0.028632
Healthy life expectancy at birth	0.509504	6.567243e-01	0.357101	1.000000	0.135857	2.033799e-04	0.086488	0.049299	0.015442	0.030842
Freedom to make life choices	0.282297	1.307122e-01	0.167379	0.135857	1.000000	1.078303e-01	0.230967	0.331901	0.074209	0.158695
Generosity	0.035182	6.990402e-07	0.004866	0.000203	0.107830	1.000000e+00	0.079014	0.097067	0.007494	0.083621
Perceptions of corruption	0.184338	1.182077e-01	0.049538	0.086488	0.230967	7.901381e-02	1.000000	0.078229	0.071421	0.211273
Positive affect	0.260656	5.777275e-02	0.181959	0.049299	0.331901	9.706734e-02	0.078229	1.000000	0.107261	0.014183
Negative affect	0.101069	5.102042e-02	0.174948	0.015442	0.074209	7.493653e-03	0.071421	0.107261	1.000000	0.015515
Confidence in national government	0.006540	3.621547e-02	0.028632	0.030842	0.158695	8.362076e-02	0.211273	0.014183	0.015515	1.000000

Figure 1: Correlation Coefficients

I noticed a few things from these values. First, I saw that Life Ladder is correlated fairly strongly with LGDP, SS, and HLE. However, LL is not correlated quite as strongly with the other predictors, and its correlations with GEN and CNG are especially weak. Even Positive Affect and Negative Affect are not correlated that strongly with Life Ladder, which is strange because I assumed those correlations would be some of the strongest.

Another small thing that caught my eye was the correlation between LGDP and HLE. This is the strongest correlation out of all the variables, which makes some sense. I also noticed that the weakest correlation by far is between LGDP and GEN, which makes sense because the values of Generosity were intentionally corrected for differences in GDP.

Next, I made boxplots for all ten variables, shown in Figure 2 on the next page. Because six of the variables strictly range from zero to one, I plotted these on the same axes, and each of the other four on their own axes. These plots showed me that the variables SS, FLC, PC, Positive Affect, and HLE are all concentrated near the high end of their respective ranges. They also showed that Negative Affect is concentrated near the low end of its range, and that there are slightly more negative observations of GEN than positive ones. Finally, I saw that CNG, LGDP, and LL all range very far, with less concentration than the other variables.

For the most part, it seems like the predictors are concentrated in favor of happiness. That is, positive predictors like Social Support are generally high, and negative predictors like Negative Affect were generally low. While the prevalence of negative values for Generosity goes against this somewhat, the far more notable exception is Perceptions of Corruption. I anticipated this being a negative predictor of happiness, and yet observations of this predictor are concentrated rather high.

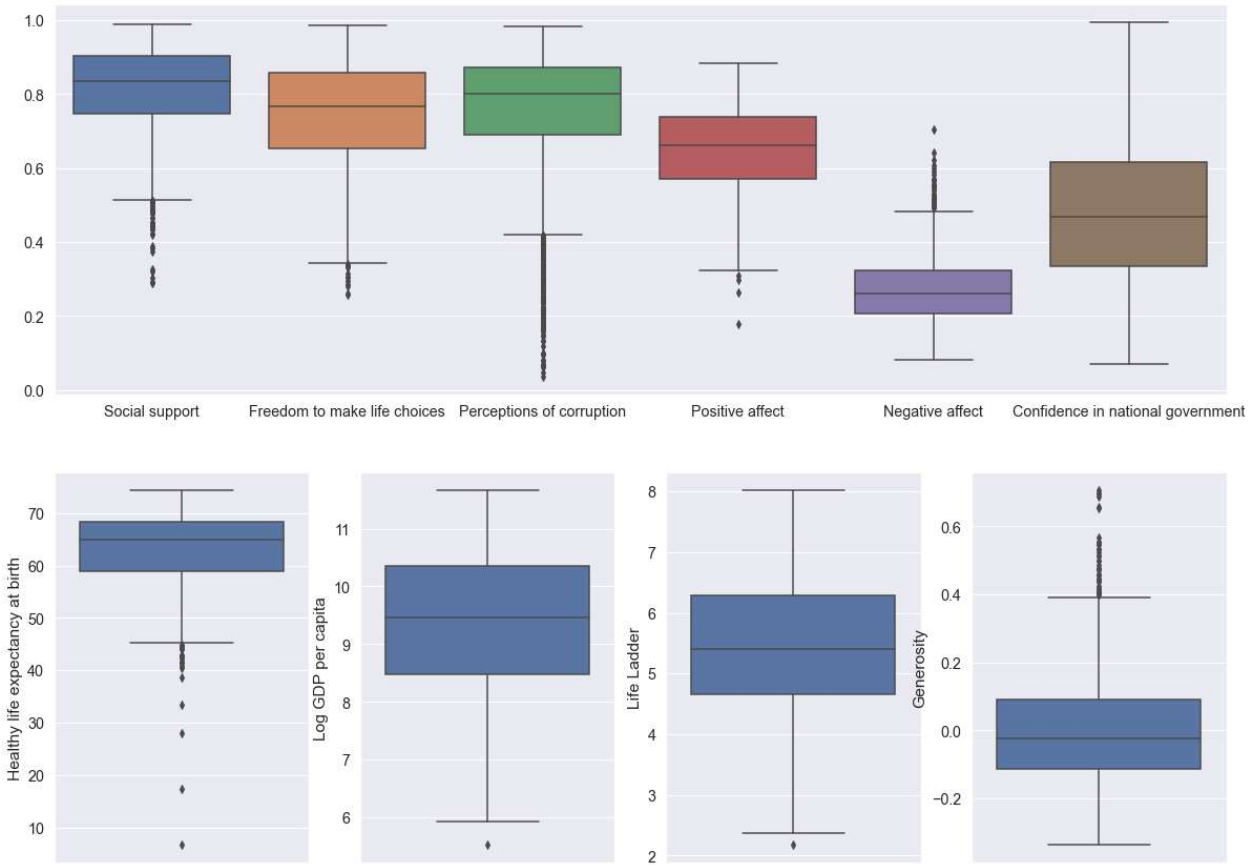

Figure 2: Boxplots

Finally, I created scatterplots (shown in Figure 3 on the next page) relating LL to each of the nine other variables. What these scatterplots show largely reflects what I learned from the r-squared values. Some kind of linear relationship can be detected in most cases, even though a few of these relationships appear relatively weak, like with Positive and Negative Affect. A few of the variables, such as SS, seem to have somewhat curved relationships with LL, which later led me to attempt transformations to make those relationships more linear. Furthermore, GEN and PC both hardly seem to have any relationship at all to LL, which matches the low r-squared values I observed for those earlier.

Something that caught my eye was the odd shape of the scatterplot relating PC to LL. It seems like the concentration on the high end of the scale that I observed from the boxplots largely belongs to countries with a Life Ladder score of 7 and below. For these countries, there seems to be no trend between happiness and perceptions of corruption. However, as soon as we look at countries with Life Ladder scores between 7 and 8, we suddenly see much lower perceptions of corruption, as well as what looks like a negative trend. These two different behaviors made me wonder if another phenomenon was at work separating the two apparent groups. I explored this more later into the project.

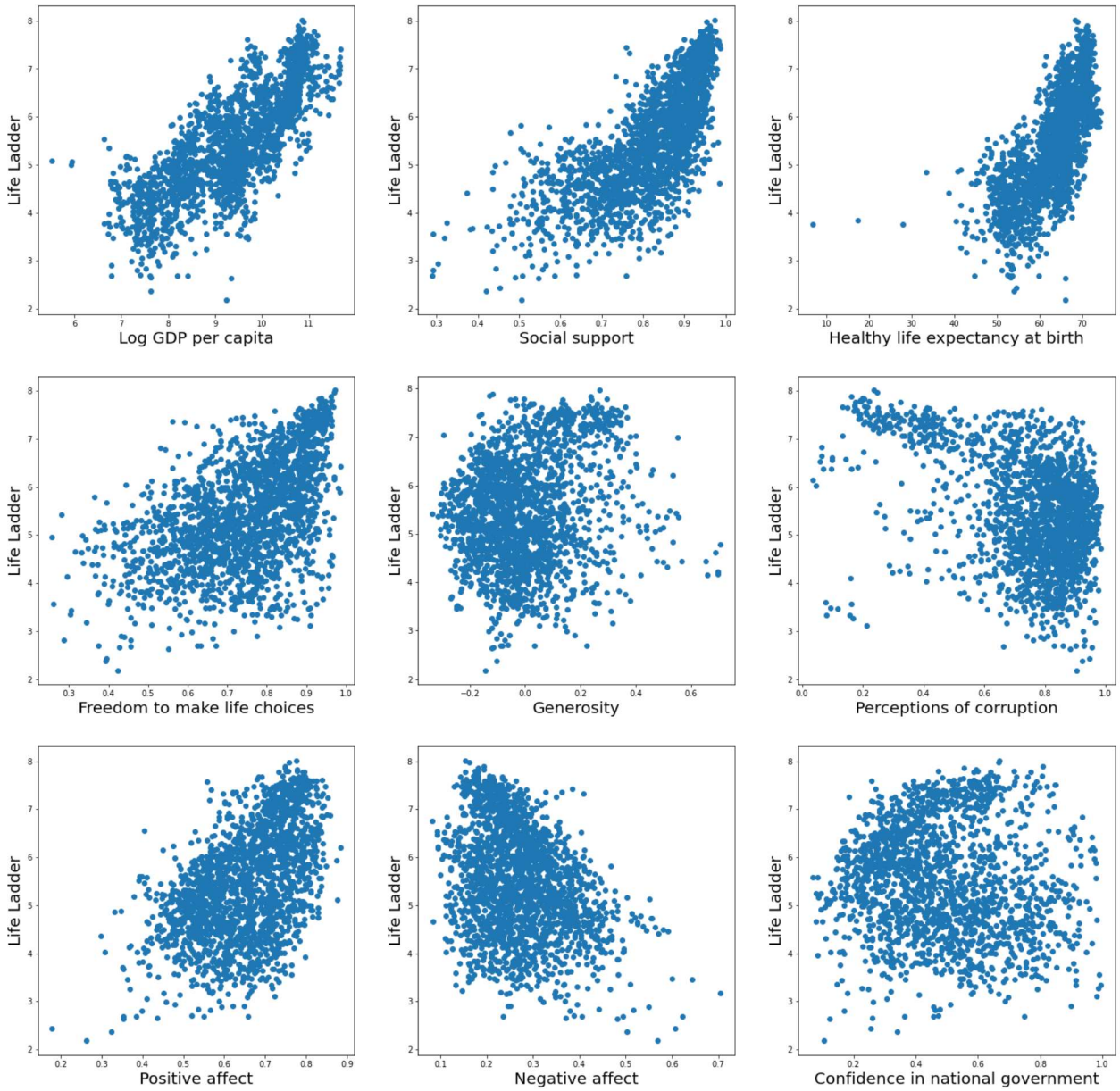

Figure 3: Scatterplots

Model Selection

The primary algorithm that I used to model my data is the linear regression algorithm from Scikit Learn, because this is what best fits the purposes of my project. For one thing, linear regression is a supervised algorithm, which was important for my labeled data. Also, I did not use a classification algorithm like logistic regression or a support vector machine because it was not my goal to sort countries into categories. I wanted a model that could predict a country's exact happiness level as closely as possible based on the given predictors, and linear regression was the best algorithm I knew of for this kind of task.

I initially chose to build my model to predict a country's Life Ladder score based on seven predictors: LGDP, SS, HLE, FMC, GEN, PS, and CNG. I chose not to use Positive and Negative Affect, because these seemed like phenomena that do not cause happiness, but rather are simply correlated with it. I was interested in identifying the biggest causes of national happiness, so I left them out of the model so that they would not distort the importance of the other seven, more causational variables.

When I built this 7-predictor linear model, 10-fold cross validation showed that its average root-mean-squared error on the testing set was approximately 0.5496. I then ran a nested for-loop to build models with all 128 possible subsets of these predictors, but all of these models turned out to have larger RMSEs than the full model, so I went forward using the full model.

In an attempt to improve the model, I took another look at the predictors that, in my exploratory data analysis, had appeared to have a curved relationship with LL. These were SS, HLE, FMC, and PC, and for each one I chose the transformation that made the relationship appear most linear. I transformed SS by raising 10 to the power of each value, and I transformed the other three features with an exponential (raising *e* to the power of each value). These transformations were meant to spread larger values farther apart from one another in the x-direction, thus lessening the curve.

I then created new models using these transformed variables in place of their raw counterparts. Ultimately, I found that the transformed versions of SS and PC improved the model, while the transformed versions of HLE and FMC did not, so I kept the SS and PC transformations and discarded the others. This new model had a 10-fold cross validated RMSE of approximately 0.5447, which was less than the previous RMSE of 0.5496.

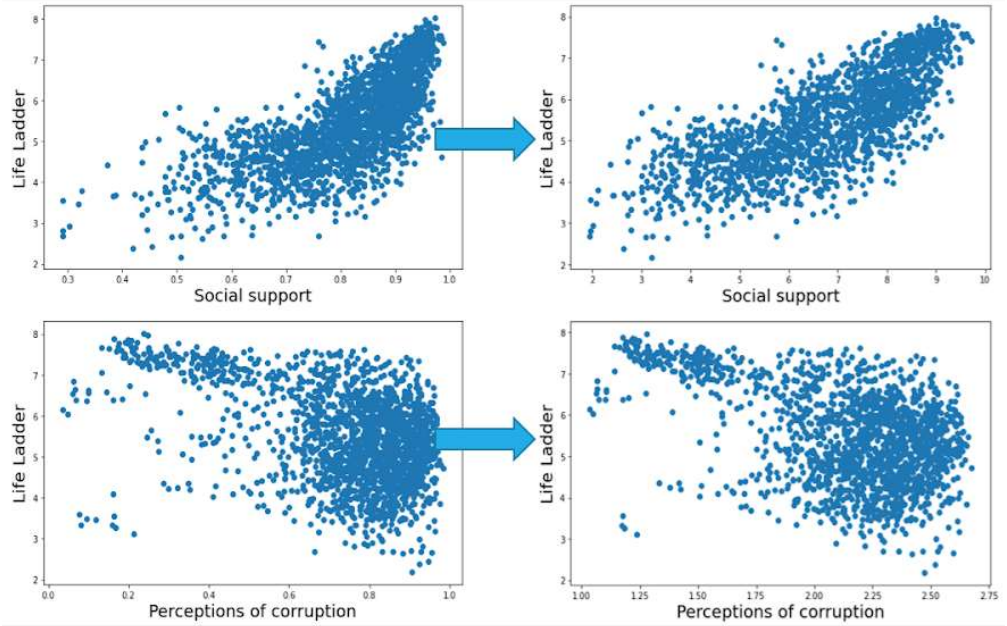

Figure 4: Variable Transformations

Next, I looked back at the variable Perceptions of Corruption, which I noticed during my exploratory data analysis had a strangely shaped relationship with LL. Because the points on that scatterplot had appeared to exhibit two clusters, I used an unsupervised k-means algorithm to identify these clusters. I set the number of clusters to two, because I could make out two clusters with my eye. The algorithm ran on a one-dimensional dataset containing only PC (this was the transformed version of PC, keeping consistent with the model). I would have included LL in the algorithm to help it better identify the clusters I saw with my eyes, but if I had done this, then any information I got from the clustering would have held implicit information about LL. This would have made this information not fair for use in the predictive model.

The clusters chosen by the algorithm (shown in Figure 5 below) match the clusters I envisioned decently closely. With this information, I added two brand new predictors to the larger dataset. The first of these was an indicator variable reflecting which of the two clusters the k-means algorithm predicted each country would fall in based on its value for PC. The second predictor I added was an interaction term between PC and the indicator variable, representing their product. These two new variables essentially allowed the model to draw two different regression lines, reflecting the two seemingly-different behaviors of the two clusters. The indicator variable allowed for lines with different intercepts, and the interaction term let the lines have different slopes as well. This is illustrated in Figure 6 on the next page. Using these two additional variables, I was able to create an even better model than I had without them.

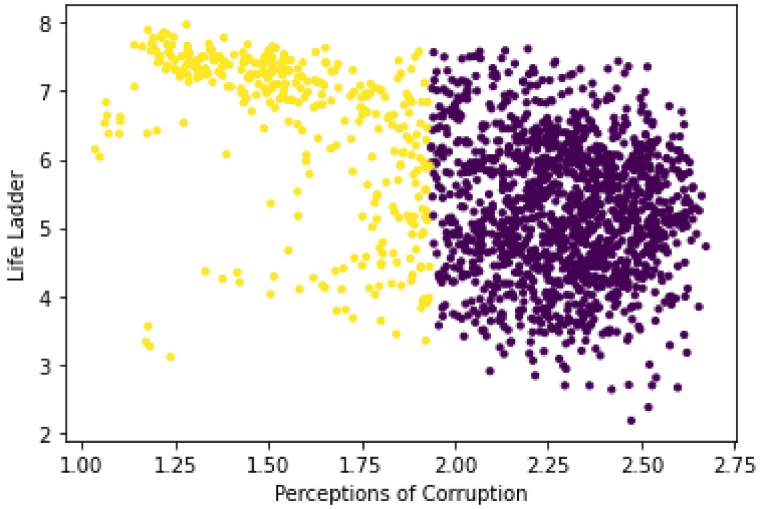

Figure 5: PC Clusters

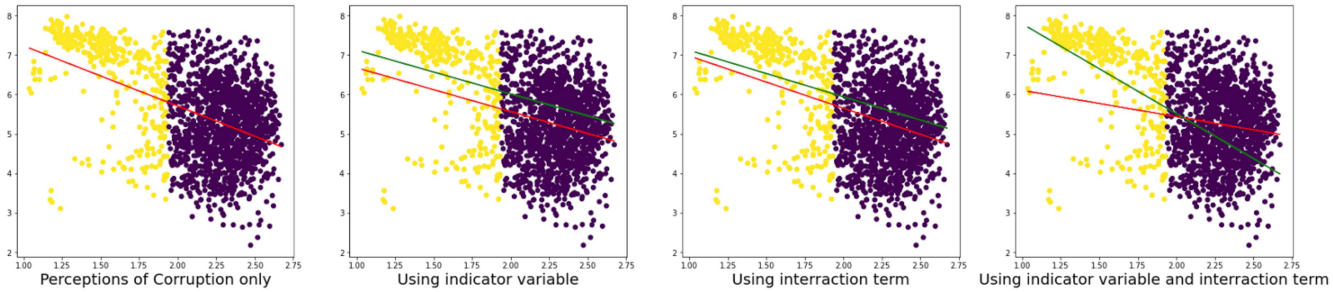

Figure 6: Regression Lines Using New Variables

Results and Evaluation

Now with nine predictors in my dataset, including the two I derived from k-means, I ran another nested for-loop to build models using all 512 subsets of these predictors, and I ran 10-fold cross validation on each of these models. The best model found by this loop was not the full nine-predictor model, but rather an eight-predictor model. The first seven of these predictors were LGDP, SS, HLE, FMC, GEN, PC, and CNG, and the eighth predictor was the indicator variable for which cluster a country falls in.

Out of every possible model, this eight-predictor model yielded the highest 10-fold cross validated r-squared score: approximately 0.7719. This model also yielded the lowest root-mean-squared-error score: approximately 0.5407. This means that, by both metrics, these eight predictors make a better model than any other subset of the nine predictors, including the full nine-predictor model as well as the seven-predictor model that did not use the predictors obtained from k-means. That is why I settled on this eight-predictor model as the ideal model to predict a country's happiness.

To demonstrate the goodness of fit of this model, Figure 7 on the next page shows the nine scatterplots from Figure 3 again. The blue dots represent the true data, and the red dots represent the predictions made by the optimal model when trained on all that data. It can be seen that the predictions match the true values very closely.

Finally, in order to study which predictors were most and least important to the model, I compared models with smaller subsets of predictors with each other. That is to say, I found the best seven-predictor model, as was the best six-predictor model and the best five-predictor model, and so on. As the number of available predictors shrinks, the first predictors to disappear from the optimal models are GEN, HLE, and the predictors obtained from k-means, which suggests that these predictors are the least necessary and add the least predictive power to the model. The predictors obtained from k-means are the first to go, because even though using one or both of them in the model seems to improve it, this improvement is incredibly small. Generosity is the next least important, which makes senses because the EDA revealed that its correlation with Life Ladder is very low. Healthy Life Expectancy at Birth is the least important after that because, even though it was found to have a rather high correlation with Life Ladder, it also had high correlations with LGDP and SS, which are both correlated even more strongly with

Life Ladder. It is likely that the same variability explained by HLE is explained better by LGDP and SS together, so their inclusion in the model makes HLE unnecessary.

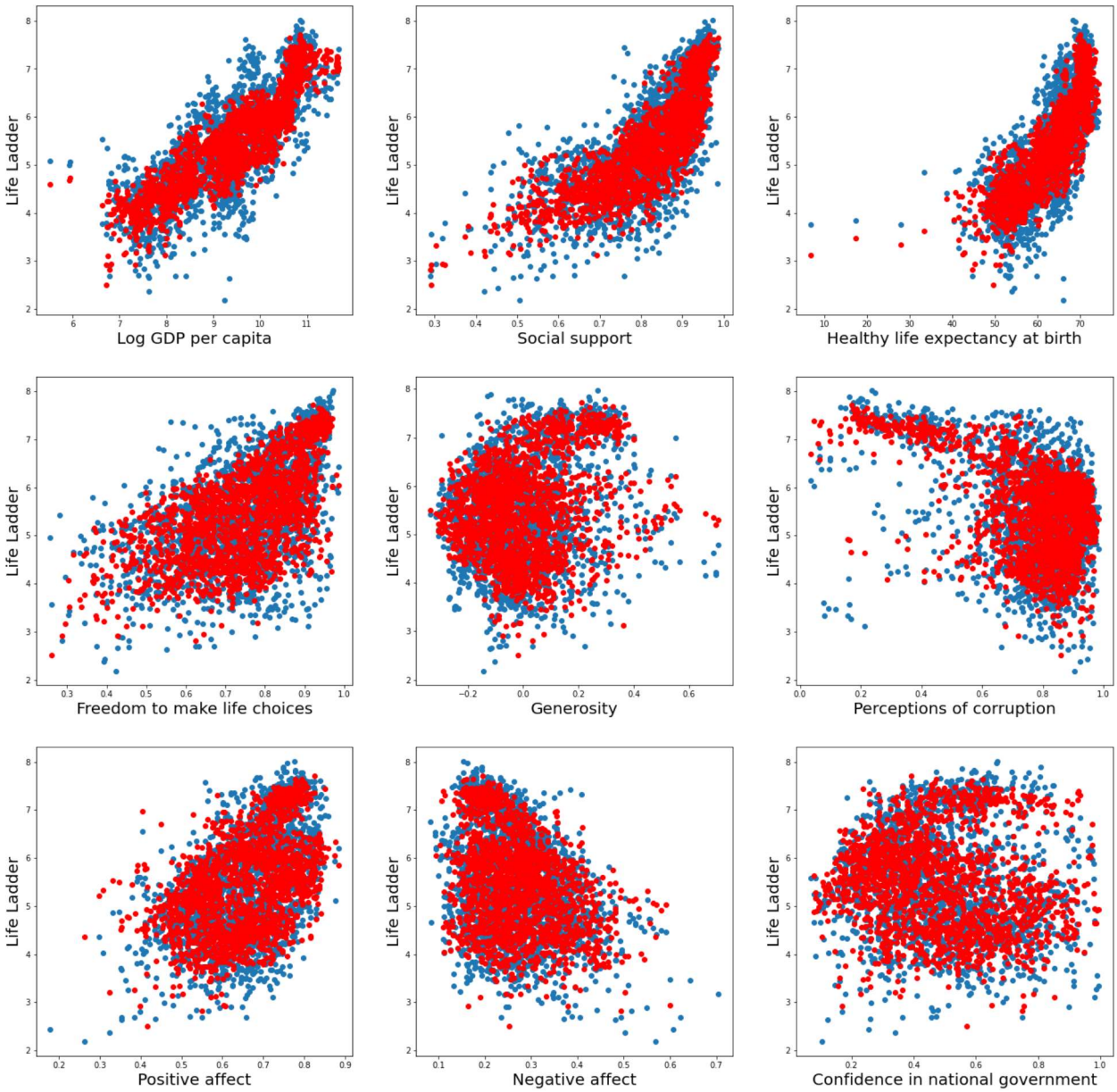

Figure 7: True Data & Predictions

Conversely, looking at the last predictors to be discarded revealed that the most necessary predictor of happiness is LGDP, followed by FMC, followed by SS. This makes sense because, not counting Healthy Life Expectancy at Birth, these are the three predictors most closely correlated with Life ladder, as found during the EDA. However, it is interesting that both Log GDP and Social Support turned out to be so important, given how closely correlated they are

with each other. As the number of available predictors shrunk, it seems like one should have made the other unnecessary, but instead the most optimal models tend to use both of them.

Through this work, I succeeded in my original goal. I created a quite accurate model for predicting a country's happiness levels, and I also identified which predictors of happiness are most and least important.

Ethics

The primary ethical concern that I saw in this kind of study was the reflections it could potentially have on different people groups around the world. To make up an example, let's say the data revealed that Asian countries are generally less happy than other countries. If researchers were not careful with how this data is handled and presented, people might infer a causal relationship between race and happiness. Similar negative associations could arise from other variables in the data too; to make up another example, say countries in Central America were found to be especially ungenerous. Of course, there may be some cases where a causal relationship does exist, like if countries where a certain religion is practiced were found to be lower on freedom to make life choices because of restrictions imposed by their faith. If this data were not handled carefully and honestly, people could start treating this religious group unfairly and blaming them for the levels of well-being in those countries, despite the evidence of that being nowhere near conclusive.

My project's solution to this ethical concern was to remove the country labels from all of the data. I wanted to make discoveries from my data without any consideration of what part of the world they come from, so that any patterns I discovered would be free from bias. If I had ended up using geographical data after all, I would have kept these concerns in mind and been considerate of what conclusions I drew from my findings and what conclusions I didn't have enough evidence to make.

Another small concern was that, if a particular predictor was found to correlate strongly with a happiness, countries interested in improvement should be careful to implement that guidance wisely. For example, if Social Support had been found to be most conducive to happiness, making social support mandatory without careful consideration of the consequences may bring about new problems for certain people, and not improve lives in the country in general.

Future Work

If I had the time, there are many ways that I could go further with this exploration into how to predict the happiness levels of a country. For example, I could take another look at the variables Positive Affect and Negative Affect. I did not use these as predictors in my model because they did not strike me as causes of well-being, but rather separate effects similar to well-being that potentially had same causes. If I wanted, then, I could redo my exploration,

except using Positive Affect or Negative Affect as the target variable rather than Life Ladder. This could allow me to see how well the same predictors can predict variables related to happiness, which predictors do this best, and whether these are the same predictors that are the best at predicting happiness.

Another thing I could do would be to reconsider my usage of the variable Confidence in National Government. As I mentioned earlier, I could not find good documentation of what this variable represents or how it was collected, so it may not be wise to draw solid conclusions from a model that uses it. Because I created models for every subset of the nine predictors I used, I could go back and look at models that do not include this variable and compare them to models that do. This would let me assess what degree of value the variable adds to the model and whether I would lose anything by choosing to discard it.

Finally, if I had much more time, I could study broader questions that would require much more research and data. For example, I could attempt to answer why the graph of PC versus LL has the odd shape that I mentioned previously. Perhaps this could lead me to discover some underlying phenomenon or structure to the data that explains more about the real causes of countrywide happiness. There is also much more research I could do into what factors truly contribute the most to happiness, because this study only answers those questions partway. I know which factors are the strongest *predictors* of happiness, but I do not know whether those factors can be called the causes themselves. The correlation of several factors with one another could indicate a deeper cause underlying all of them. Also, I am not sure whether the model-importances outlined above are the best representation of which factors contribute most to happiness, or whether the correlations with happiness found through the EDA represent this better. Answering this question would require much more exploration and thought than I had time to do for this project.

References

Helliwell, John et al. “World Happiness Report.” *World Happiness Report*, https://worldhappiness.report/ed/2022/. Accessed 6 December, 2022.

Helliwell, John et al. “Statistical Appendix for ‘Happiness, benevolence, and trust during COVID-19 and beyond,’ Chapter 2 of World Happiness Report 2022.” *World Happiness Report*, March 9, 2022, https://happiness-report.s3.amazonaws.com/2022/Appendix_1_StatiscalAppendix_Ch2.pdf. Accessed 6 December, 2022.

“World Values Survey.” *World Values Survey*, https://www.worldvaluessurvey.org/wvs.jsp. Accessed 6 December, 2022.

Stevenson, Betsey and Justin Wolfers. “Economic growth and subjective well-being: reassessing the Easterlin paradox.” *CESifo Working Paper*, No. 2394, September 2008.

Data Source

World Happiness Report 2022 raw data: Data for Table 2.1