What Makes People Happy? An Analysis of Socio-Economic Factors that Impact Happiness Around the Globe

Brandon T. Pugh

October 14, 2020

1 Abstract

The United Nations, in recent years, have had a hard time determining what actually influences the level of happiness in countries around the world. This is an issue since the main goal of the UN is to improve the living conditions and happiness of the people in the world, but they cannot create the policy to do so if they don't know what influences world happiness. This analysis was able to identify several key socio-economic factors that have an impact on a nation's happiness. The UN can then use the knoweldge from these factors to create policy that can improve the lives of people all across the world.

2 Introduction

What is happiness? Many have tried to answer this question, and no one really has found a clear, definitive answer to this. The World Happiness Organization[5] is one group that has tried to answer this, and have spent years gathering data of different socio-economic factors to compile, what they call, the ”life ladder”. This is how they measure the happiness of a country, and contains a value from 0 to 10, with 0 being the lowest value of happiness and 10 being the highest. The closer a country's life ladder value, the happier it's citizens are. The purpose of this analysis is to determine what socio-economic factors have an impact on a nation's happiness, and use that information to find out which nations are doing the worse at those significant factors so that the UN can direct policy to try to improve the situation in those countries.

3 Data

3.1 Data Overview

The data that will be analyzed in this report come from the World Happiness Report, which gives reports on the happiness of different countries from around the world for each year. For how the data was actually collected, see [5]. This provides a brief explanation of the data collection process and frequently asked questions about the data.

The initial data set contains 1,848 observations across 26 different variables. The way that the data is structured is such that each unique combination of country and year has a corresponding row in the data. See Table 1 for how this data is initially structured. The variables of this data set along with their descriptions can be found in Table 2. For the remainder of the paper, I will be referring to the variable shorthand that can be found in Table 2.

3.2 Data Preprocessing and Cleaning

The program that the data was analyzed in was R version 4.0.2. After reading in the data and looking at the data summary, there are several key takeaways that can be realized from this initial look at the data. One, is that the years that the data was taken was between 2005 and 2019, but not every single country had data recorded for every single year. There are several countries that only had data for four or five of

Country.Name	Year
Afghanistan	2008
Afghanistan	2009
Afghanistan	2010
...	...
Albania	2007
Albania	2008

Table 1: Initial Data Structure in Happiness Data

Name	Shorthand	Description
Country.Name	Country	The country
Year	Year	Year where the data was collected
Life.Ladder	Happiness Level	Level of happiness
Log.GDP.per.Capita	LogGDP	Level of GDP per capita for the country
Social.Support	Social Support	Percent of people that can get help
Healthy.life.expectancy.at.birth	Life Expectancy	Average length of life
Generosity	Generosity	Percentage people that donated
Perceptions.of.Corruption	Corruption	How corrupt the government is
Positive.Effect	Positive	How often someone feels happy in a day
Negative.Effect	Negative	How often someone feels negative in a day
Confidence.in.national.government	Confidence	Percentage of people confident in their gov
Democratic.Quality	Democratic	How democratic a country is
Delivery.Quality	Delivery	How democratic a country is
Standard.deviation.of.ladder.by.country.year	SdofHappiness	measure of the inequality of happiness
Standard.deviation.mean.of.ladder.by.country.year	meanofSd	mean of the inequality of happiness
GINIWorldBank	GINIWorldBank	A measure of income for the nation
GINIHousehold	GINIHousehold	A measure of income for the household
Most.people.can.be.trusted	Trust [With Years]	Trust index for certain years, 7 different var

Table 2: Variable Descriptions

the years. Another is that there is a vast amount of missing data, with several variables containing over 50 percent of its observations missing. This was found by using the Data Explorer package in R [1].

Another key observation in this initial view of the data is that several of the variables, such as confidence, have values that range between 0 and 1. This means that values for these such variables should be interpreted as percentages. For example, if a country had a confidence value of .6785, it should be interpreted as 65.85 percent of people in that nation are confident in their government.

3.2.1 Missing Data

The first step that was taken in cleaning this data for analysis was checking the level of missing data that each variable. If a variable had more than 50 percent of it's data missing, it was dropped from the data set using the select function found in the dplyr package. [4]. The reasoning behind this decision is that there would be no way to input the data manually from each year, and any attempt to do so would lead to serious skewing in the results of this analysis. The variables that were dropped from the data set were all of the trust variables (regardless of year grouping), and a variable that estimated the GINIWorldBank variable. In total, eight variables were dropped.

3.2.2 Variable Correlation

The next step taken in the Data Preprocessing is checking the correlation of the remaining variables. What a correlation value tells us is how similar two variables are to each other. The closer that a correlation value is to either 1 or -1, the more information that is the same between the two variables. After finding the correlation values and constructing a correlation plot using the corrplot package [3], it was found that both

Democratic and Delivery had a correlation of over .8, meaning that they told the same information. So the Delivery variable was dropped from the data set, since Democratic is easier to interpret than Delivery.

3.3 Data Aggregation

After cleaning the data, the data has 1848 observations based on unique combinations of country and year, structured in Table 1. Now this particular analysis is only interested on what socio-economic factors affect a country's happiness level, not analyzing happiness over time. To deal with this issue, the data was aggregated on the country level, and taking the average of the remaining variables across the years that each country had data recorded for. Now we have 166 different observations, one for each country in the data set, with the variables now being an average for that country across all years. An illustration of how this data is now structured can be found in Table 3.

Country.Name	...
Afghanistan	...
Albania	...
Algeria	...
Angola	...

Table 3: Aggregated Structure for Happiness Data

3.3.1 Missing Data in the Aggregate

After aggregating the data up to the country level, the next step is to check the newly aggregated data for missing data. Again the Data Explorer package to do so. [1]. After checking this, there are a few instances of missing data in the aggregated data. This means for that particular variable, the country had no recorded observations for that variable.

Now this issue of missing data can be handled in a easy way. Since there are only a few instances of no recorded data for each variable in the data set, the mean of the variable can be inserted as a place holder without affecting the future analysis.

4 Methodology

4.1 The Model

Now that the data set is ready for analysis, the question one has to ask is how should the analysis should be done? The answer to this question is a multiple linear regression[2]. The reason for this is the question that this analysis is trying to answer: What affects happiness in countries around the world? This question is explanatory in nature, and a multiple linear regression is more than sufficient enough to answer this question.

The variables that are included in the multiple linear regression are LogGDP, Social Support, Life Expectancy, Generosity, Corruption, Positive, Negative, Democratic, GINIWorldBank, and GINIHousehold. The reason why Country.name was excluded from the model is that it would perfectly predict happiness for each country if it was included, and would be useless in this analysis. For the sdofHappiness and meanofHappiness, since they are different ways of measuring a countries happiness, they were also not included in the model.

5 Results

5.1 Model Results

After conducting the analysis, the results of the model can be found in Table 4. The variables that are significant in this model (meaning that they have some sort of impact on the happiness in a nation) are as followed: LogGDP, Social Support, Life Expectancy, Corruption, Positive,Cconfidence, and GINIHousehold.

The adjusted R-squared value of this model is 84 percent. This means that 84 percent of the variation of the happiness value is captured by this model.

The variables that have a positive effect on a country's happiness are: LogGDP, Social Support, Freedom, and Positive. The variables that have a negative effect on a country's happiness are Corruption, Confidence, and Household.

	Model 1
(Intercept)	-0.29
	(0.78)
LogGDP	0.24^{***}
	(0.06)
Social Support	1.56^{**}
	(0.49)
Life Expectancy	0.02^{*}
	(0.01)
Freedom	1.15^{**}
	(0.43)
Generosity	0.37
	(0.28)
Corruption	-1.11^{***}
	(0.27)
Positive	2.61^{***}
	(0.62)
Negative	1.20
	(0.62)
Confidence	-1.14^{***}
	(0.27)
Democratic	0.05
	(0.06)
GINIWorldBank	-0.60
	(0.63)
GINIHousehold	-1.01^{*}
	(0.50)
R^2	0.86
Adj. R^2	0.84
Num. obs.	166

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Regression Model

6 Discussion

6.1 Interpretation

Most of the variables impact happiness in ways that one might expect. For instance, it is expected that the more freedom that one nation's citizens has, the more happy that country becomes. Or when there is an increased level of corruption in a country, the less happy one's citizens become. There is one variable that doesn't behave as one might expect, and that is confidence.

According to the model results, the more confidence that citizens have in their government, the less happy they are overall. This is a surprising result since it would be expected that the more confidence a nation's citizens are in their government, the more happy they would be. But this is not the case. How could this possibly be? It is the opinion of the author that it could be possible that the nation's that have a higher confidence have an average education level of their citizens is less than those nations that have a lower confidence in their government. A lower education level could make a population more likely to believe in everything that their government does, even if those actions cause a detriment to the nation.

6.2 Impact of the Results

Now that the analysis of what affects world happiness is completed, how does that help the United Nations improve the levels of happiness across the world? By now knowing which socio-economic factors impact happiness, they can then see which countries do the worst for those particular factors. After seeing which countries do the worst, then they can investigate as to why their values for those variables are so low, and then implement policy that can help increase the quality of life for the citizens of that country.

6.3 Limitations

There are some limitations with this analysis. For one, there are plenty of other factors that could potentially impact a nation's happiness level that were not included in this analysis. Other socio-economic factors such as access to clean water, access to plumbing, average food consumption and others were not considered. Also physical factors such as the average rainfall and temperature were not considered. A good continuation is to see whether or not these other factors can also contribute to a nation's happiness score.

Another limitation of this analysis is that due to time constraints, visualizations detailing which countries have the highest or lowest values for each of the significant variables. These would have been able to detail which countries struggle with the indicators of happiness, and then the UN can create policy that can target these nations to improve their happiness. A future analysis should include these visuals

7 Conclusion

In conclusion, this analysis was able to answer the question of what affects world happiness. By completing a multiple linear regression, it was determined that GDP, Social Support, Life Expectancy, Freedom, Corruption, Positive, Confidence, and Household. The United Nations can then use this information to then identify which nations struggle with these factors, and build policy to help improve the happiness in those nations. Future analysis could include more factors such as access to clean water, food, and plumbing, as well as physical factors such as weather and elevation. This could provide more insight into additional factors of happiness.

References

[1] Boxuan Cui. *DataExplorer: Automate Data Exploration and Treatment.* R package version 0.8.1. 2020. URL: `https://CRAN.R-project.org/package=DataExplorer`.

[2] *Multiple Linear Regression.* URL: `https://www.investopedia.com/terms/m/mlr.asp`.

[3] Taiyun Wei and Viliam Simko. *R package "corrplot": Visualization of a Correlation Matrix.* (Version 0.84). 2017. URL: `https://github.com/taiyun/corrplot`.

[4] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation.* R package version 1.0.2. 2020. URL: `https://CRAN.R-project.org/package=dplyr`.

[5] *World Happiness Report.* URL: `https://worldhappiness.report/ed/2020/`.