World Happiness Survey Analysis

Christian Berardi

October 2018

Background

Following the adoption of UN Resolution 65/309 in 2011, the UN has released, with the exception of 2014, yearly reports on global happiness. [1] These Happiness Reports are a result of global surveys conducted under the auspicious of the UN. [2] A number of different variables, along with happiness are included in the World Happiness Report. [3] For the purpose of this analysis, the following variables will be used as predictors for a country's happiness: Economy, Family, Health, Freedom, Generosity, Trust.

Variables

Each variable will now be defined. [4]

Happiness

The target, happiness, is the national average of ranking, from 0 to 10, of how well a person believes their life is going, from the worst possible, to the best possible.

Economy

Economy is a measure of national GDP in PPP per capita in constant 2011 dollars in August of 2016.

Family

Family is the average of an indicator measuring whether or not the respondent has someone, a friend or family member, who would help them in a time of personal crisis.

Health

Health is a measure of healthy life expectancy at birth, not simply life expectancy at birth.

Freedom

Freedom is the average of an indicator of whether or not the respondent is satisfied with their freedom to choose what to do in life.

Generosity

Generosity is a measure of the frequency of charitable donation per month scaled with GDP per capita.

[1] Happiness: towards a holistic approach to development: resolution adopted by the General Assembly, http://repository.un.org/handle/11176/291712

[2] World Happiness Report, http://worldhappiness.report/faq/

[3] World Happiness Report 2017, http://worldhappiness.report/ed/2017/

[4] World Happiness Report Statistical Appendix, https://s3.amazonaws.com/happiness-report/2017/StatisticalAppendixWHR2017.pdf

Trust

Trust is a measure of corruption perception, the average between two indicators: existence of widespread public corruption and existence of widespread private corruption.

Analysis

In order to better understand the relationship between the different predictors, the data will first be explored. This exploration will look for any problems with the data itself, either missing or questionable values. Correlations between the predictors will be calculated to determine if multicolinearity will be an issue with this data.

Following that, the relationships between the predictors and the relationships between the nations investigated in the survey, will be measured using various unstructured learning techniques. Principle Component Analysis (PCA), will be done to better understand the relationships between the different predictors, and to serve as predictors if multicolinearity proves intractable. A Bi-Plot will then be created to identify outlier nations that require special consideration and interpretation. Following that, Hierarchical Clustering (HC) will be done to help to identify patterns among the nations in the survey, these patterns will also be identified through K-Means Clustering analysis.

Subsequent to unstructed learning, structured learning, namely three predictive modeling algorithms, with be done to predict happiness using the 6 predictors identified for this analysis. 3 models will be fit to the data: LASSO, Random Forest (RF), and Support Vector Machine (SVM). These three techniques will undergo hyperparameter optimization, model, validation, and finally model comparison to the select the best fit for the target. Once the optimum model is found, it will be interpreted and conclusions from the entire analysis.

Data Exploration

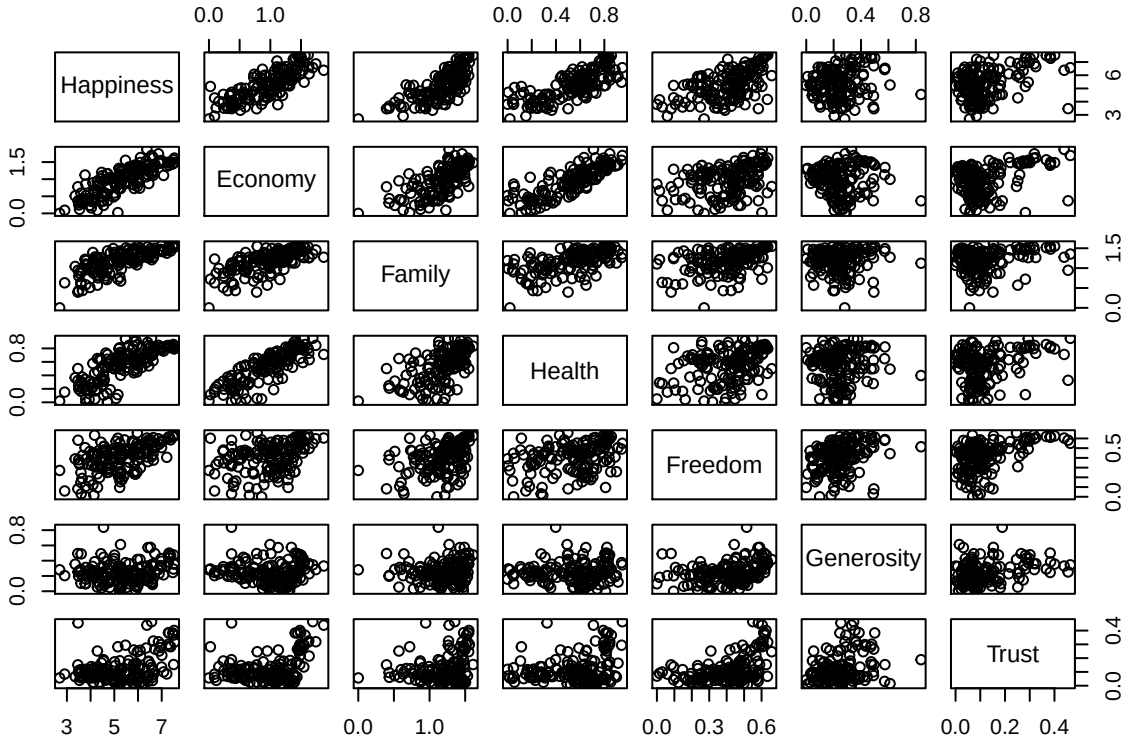
In order to better understand the relationship between the different predictors, as well as to be begin to be able to appropriately model the data, some basic statistics, as well as correlation, and the relationship between pairs of predictors will be explored.

From Table 1, the data contains 155 different nations. Looking at the summary statistics calculated for each of the 6 predictors used in the report, none have missing values, nor do they have maximum values much greater than their 3rd quartile. For this reason no work need be done on data cleaning. This is expected of a data set from Kaggle.

Table 1: Summary of Data

Country	Happiness	Economy	Family	Health	Freedom	Generosity	Trust
Afghanistan: 1	Min. :2.693	Min. :0.0000	Min. :0.000	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.00000
Albania : 1	1st Qu.:4.505	1st Qu.:0.6634	1st Qu.:1.043	1st Qu.:0.3699	1st Qu.:0.3037	1st Qu.:0.1541	1st Qu.:0.05727
Algeria : 1	Median :5.279	Median :1.0646	Median :1.254	Median :0.6060	Median :0.4375	Median :0.2315	Median :0.08985
Angola : 1	Mean :5.354	Mean :0.9847	Mean :1.189	Mean :0.5513	Mean :0.4088	Mean :0.2469	Mean :0.12312
Argentina : 1	3rd Qu.:6.101	3rd Qu.:1.3180	3rd Qu.:1.414	3rd Qu.:0.7230	3rd Qu.:0.5166	3rd Qu.:0.3238	3rd Qu.:0.15330
Armenia : 1	Max. :7.537	Max. :1.8708	Max. :1.611	Max. :0.9495	Max. :0.6582	Max. :0.8381	Max. :0.46431
(Other) :149	NA	NA	NA	NA	NA	NA	NA

Correlation Between Predictors



The plot above shows clear evidence of high correlation between multiple predictors. For this reason controlling multicolinearity will need to be a focus when modeling the data.

Table 2: Correlation between Predictors

	Happiness	Economy	Family	Health	Freedom	Generosity	Trust
Happiness	1.0000000	0.8124688	0.7527367	0.7819506	0.5701372	0.1552558	0.4290797
Economy	0.8124688	1.0000000	0.6882963	0.8430766	0.3698734	-0.0190113	0.3509441
Family	0.7527367	0.6882963	1.0000000	0.6120801	0.4249658	0.0516926	0.2318414
Health	0.7819506	0.8430766	0.6120801	1.0000000	0.3498268	0.0631915	0.2797520
Freedom	0.5701372	0.3698734	0.4249658	0.3498268	1.0000000	0.3160827	0.4991828
Generosity	0.1552558	-0.0190113	0.0516926	0.0631915	0.3160827	1.0000000	0.2941595
Trust	0.4290797	0.3509441	0.2318414	0.2797520	0.4991828	0.2941595	1.0000000

Looking further into correlation, we find the same issue when calculating the correlation between the predictors. Table 2 shows correlation between many of the predictors is large.

This correlation indicates many predictors carry the same information, which exposes the multicolinearity risk inherent in modeling this data.

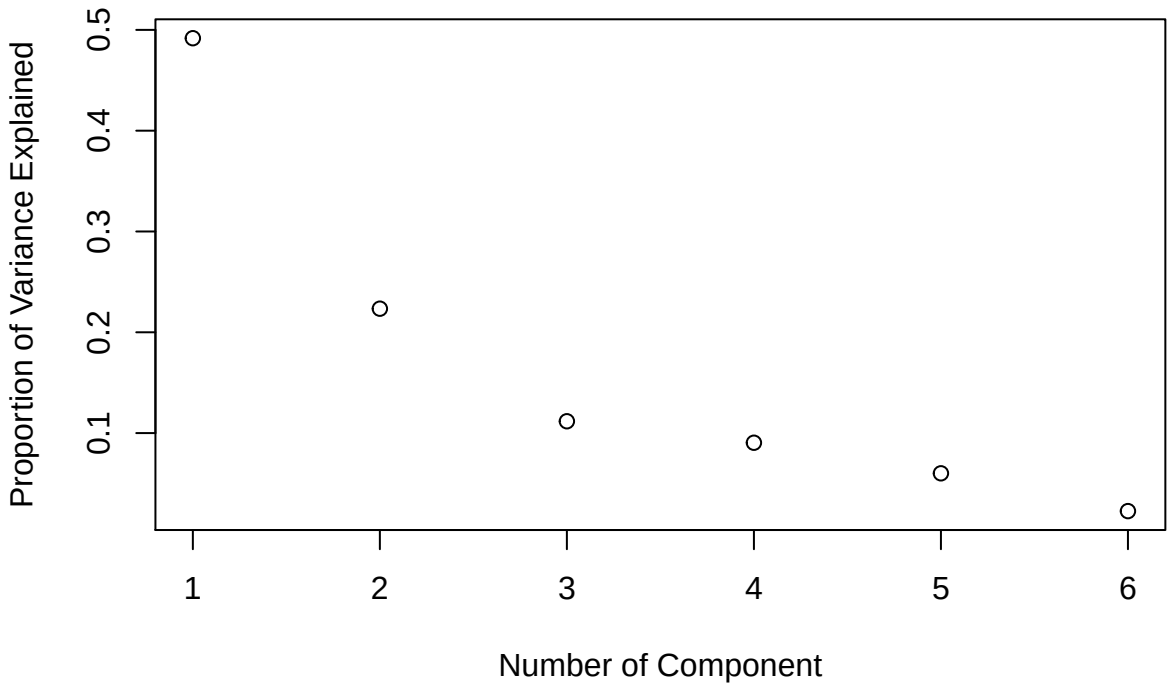
Table 3: PC 1

	x
Economy	0.5086686
Family	0.4625428
Health	0.4886628
Freedom	0.3938236
Generosity	0.1379690
Trust	0.3382111

Unstructed Learning

To obtain a better understanding of the nations and relationships between the correlated predictors, as well as to discover outliers and cluster the nations of the world, PCA, HC, K-Means Clustering were carried out and Bi-Plots were created.

Principle Component Analysis



The scree plot indicates that 3 principle components(PC) needs to be interpreted. The 1st principle component is:

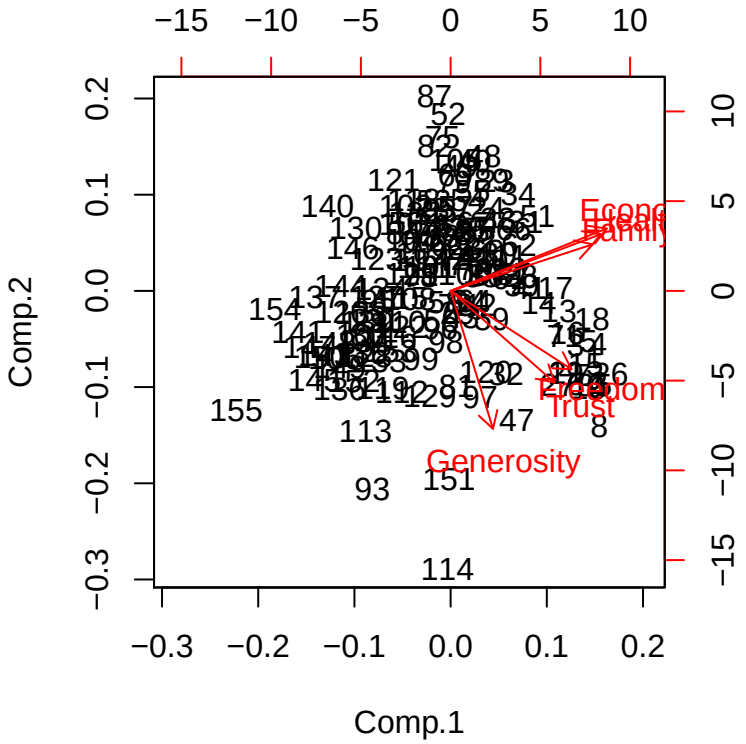
This PC indicates a weighted average of the different predictors. This gives us no special insight into the relationship between the predictors.

Table 4: PC 2 and 3

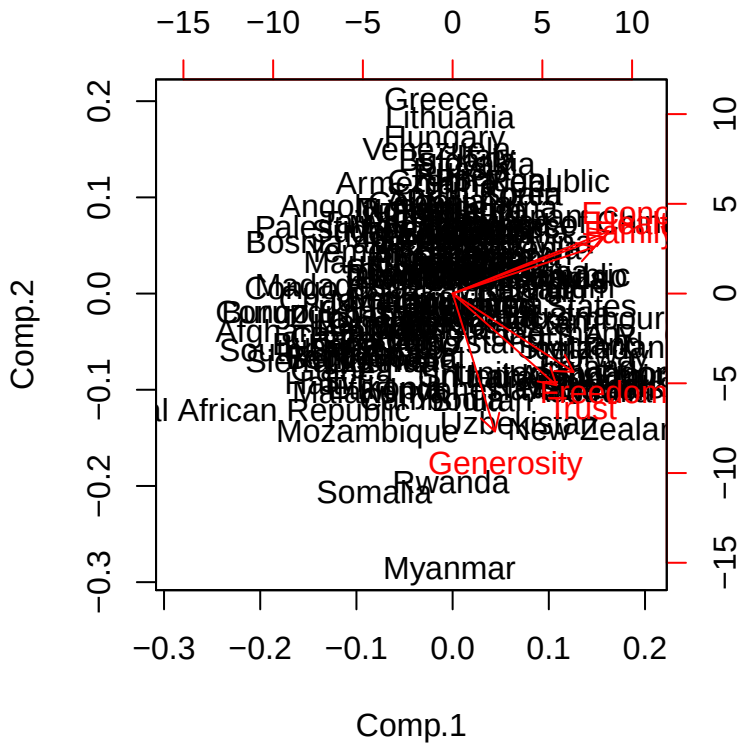
	PC2	PC3
Economy	0.3012362	0.0197682
Family	0.2250019	0.2044319
Health	0.2735209	0.1956850
Freedom	-0.3789218	-0.2171700
Generosity	-0.6663931	0.6958410
Trust	-0.4428948	-0.6230302

These two PCs are much more interesting. The 2^{nd} PC is a contrast between economy, family and health on one hand and freedom, generosity and trust on the other. This can be interpreted as suggesting that there are two classes of predictors for this data, one I will call goodness of life predictors: economy, family and health, the other goodness of government: freedom, generosity and trust. The 3^{rd} PC moves generosity from the goodness of government to goodness of life predictors. The conclusions that can be drawn form this will be considered following bi-plot analysis.

Bi-Plot



To check the conclusions from the (PCA), a bi-plot was constructed. The bi-plot results are very similar to those seen from PCA, the goodness of life predictors are clustered, as are two of the goodness of government: trust, freedom. Generosity is now found on its own, which is consistent with its change from one contrast to the other seen in PCA. MORE



A second bi-plot was constructed to determine which countries could be seen as outliers so that they could be better understood. We immediately see a number of countries which deserve further scrutiny: Myanmar, Rwanda, Somalia, Greece, Lithuania, and Taiwan (hidden by the goodness of life predictors).

Myanmar

Myanmar is ranked low in term of happiness, 114 out of 155. Its outlier nature is due to its much higher then expected generosity score given its happiness– it has the highest score. The most likely explanation for this discrepancy is the highly Therevadan Buddhist nature of the majority Bamar ethnic group who, by the definition of the predictor, must be very willing to donate to charity.

Rwanda

Rwanda is also ranked very low in terms of happiness, 151 out of 155. Its outlier nature is due to its very high trust score given its low happiness. This is most likely due to the after effects of the Rwandan genocide, and the extreme changes it brought to the Rwandan government, eliminating a large amount of the corruption endemic to the region.

Somalia

Somalia is ranked near the middle in terms of happiness, 93. However, Somalia's economy score is second to last, which would usually indicate an extremely low happiness rating. The reason for this difference is most likely the continued factional nature of the Somali state. It might also indicate a need to address the adequacy of economic indicators in Somalia, and leads to questions about their accuracy.

Greece

Greece has a middling happiness level, ranked 87, which is in sharp contrast to its generosity, ranked last, which would usually indicate a much lower happiness. This is most likely a consequence of the continued financial instability of the country resulting a lack of willingness to give to charity when an individual is already experiencing a decline in living standards.

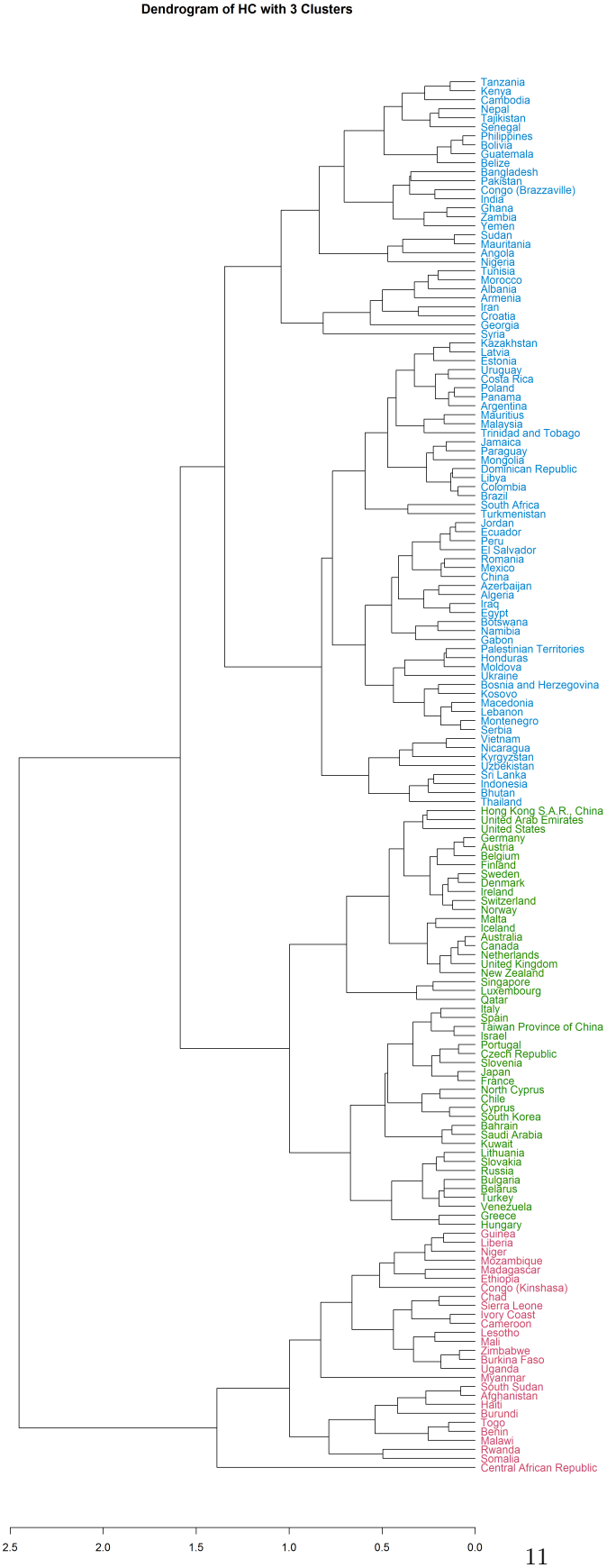
Lithuania

Lithuania is ranked in the top third of countries, which is surprising give its low generosity score, only slightly higher than Greece's. This would normally indicate a much less happy nation. Lithuania, like Greece, also recently went through financial instability. For this reason its interpretation matches that of the Hellenic State.

Taiwan

Taiwan is near the top of the rankings in terms of happiness, ranked 33. However its low freedom score would normally be associated with a much lower happiness. The reasons for this perception of a lack of ability to choose what one does in life would require further research to fully understand.

Hierarchical Clustering

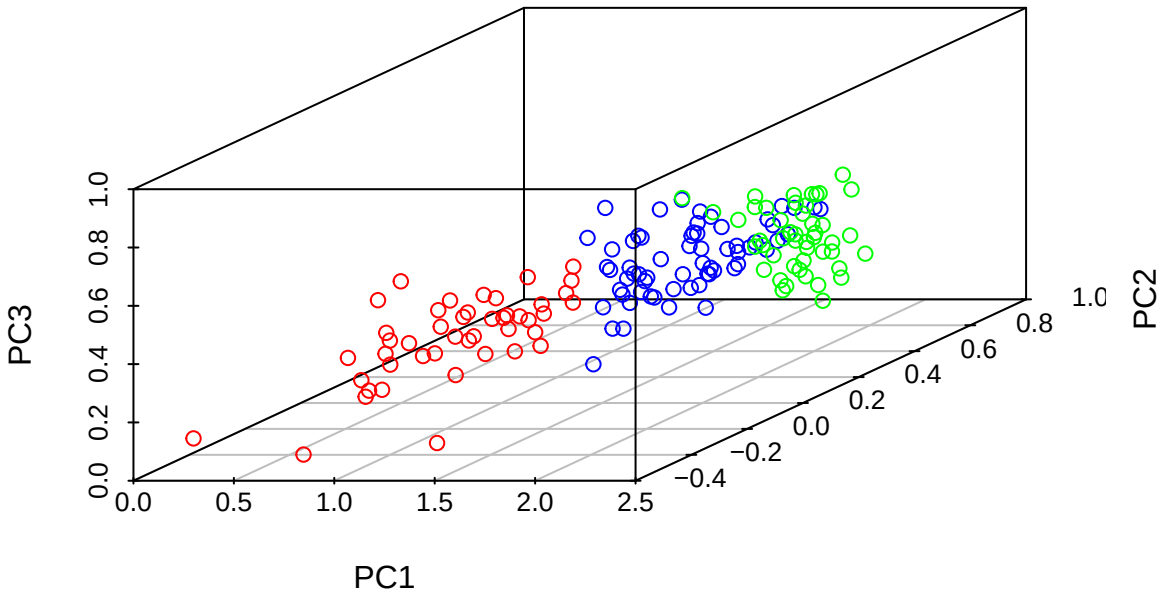


HC was done to better understand the relationships between the different countries, that is, to see if there was some way to classify the countries based on their predictor values. To that end complete clustering was done which produced the dendrogram seen on the previous page.

Interpreting the results, it is the second split that is most meaningful. This splits the countries into three groups: those with very low happiness, pink, green, the second branch, and those with middle happiness, blue.

This 3-way split corresponds, in some way, to the old Cold War, 3 World classification. However, it should be noted that the "3rd World" has greatly shrunk, while the 2nd World, previously the domain of communist regimes, has now become the most numerous. The 1st world has also increased in size, taking former 2nd and 3rd World nations.

K-Means Clustering



This division of the data into 3 parts is consistent with the results from PCA, HC also indicates there is some evidence in the data of a tripartite division in the world. For this reason, K-means clustering was done with a cluster number of 3.

The plot above shows the results of this clustering. Further visualization and interpretation of the cluster will done using Tableau.

This plot from Tableau shows a clear indication of the meaning of the 3 clusters identified above. One cluster corresponds to the highest rated countries (green), one the lowest (red), and one to everyone else (yellow). These results correspond, almost completely, with the current developed, developing and under developed nations of the world.

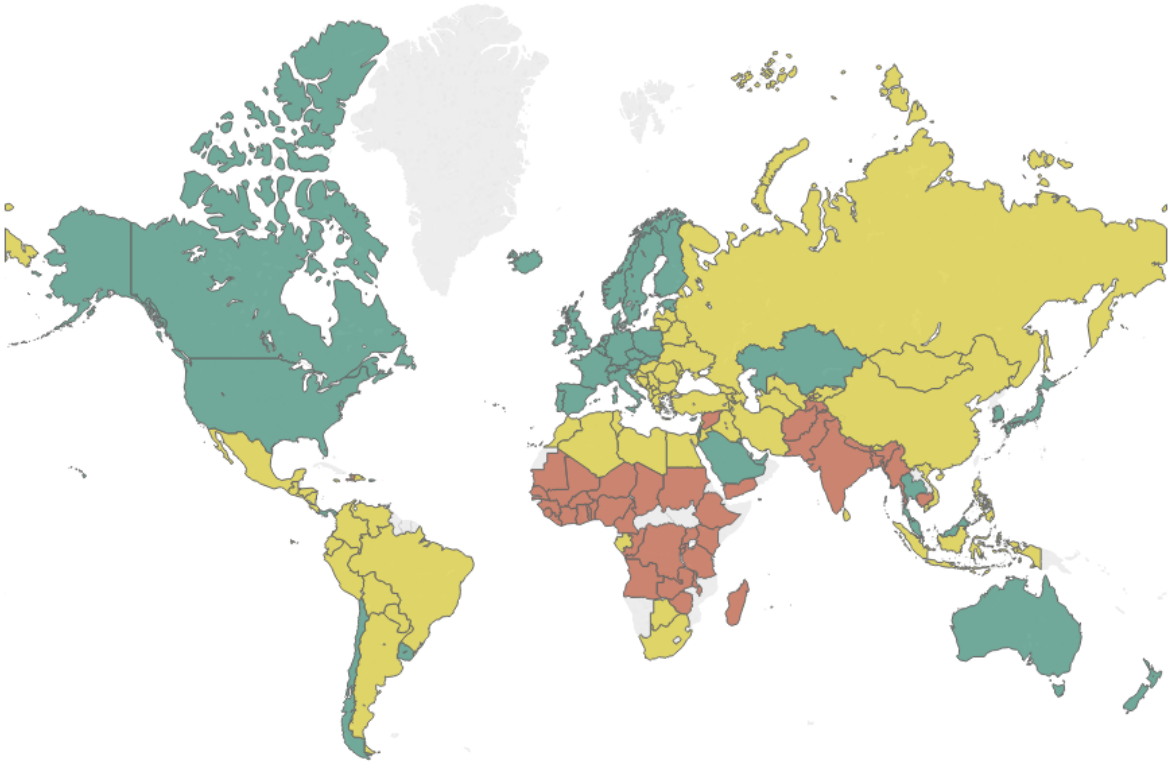


Figure 1:

Table 5: LASSO Results

	Train	Test
MSE	0.2344349	0.2721838
MSE Sd.	0.0223651	0.0545674

Structured Learning

Regression

In order to help overcome issues with the multicolinearity inherent in the data, the LASSO will be used to obtain a linear model of the data. Using 5-fold cross validation to select the optimum regularization parameter, λ, for fitting a LASSO model, the following LASSO model was fit to the data. 1000 70/30 train test splits were performed to estimate the MSE and standard deviation of the MSE for both the training and test data. The results are as follows:

Table 6: Random Forest Results

	50	100	250	500	1000	1500	2000
Mean MSE Train	0.2971371	0.2878790	0.2812844	0.2807929	0.2800743	0.2773720	0.2791625
Mean MSE Test	0.2846431	0.2820412	0.2797748	0.2789859	0.2777539	0.2809105	0.2785899
Mean MSE Sd Train	0.0304067	0.0275608	0.0277699	0.0267584	0.0267090	0.0268773	0.0264240
Mean MSE Sd Test	0.0540389	0.0524256	0.0531352	0.0509062	0.0514747	0.0522285	0.0515626
Economy	0.3800334	0.3802287	0.3797828	0.3791019	0.3793785	0.3803203	0.3823422
Family	0.2564169	0.2572916	0.2598163	0.2558010	0.2565457	0.2576721	0.2562950
Health	0.4224212	0.4221274	0.4222230	0.4250682	0.4229309	0.4279370	0.4195792
Freedom	0.1181875	0.1166902	0.1184979	0.1166354	0.1183094	0.1178243	0.1176372
Generosity	0.0141199	0.0137884	0.0144435	0.0142619	0.0142994	0.0144727	0.0146479
Trust	0.0396907	0.0396352	0.0390325	0.0398391	0.0395227	0.0397466	0.0400653
Economy Sd	0.0770720	0.0683807	0.0582149	0.0584990	0.0541186	0.0546766	0.0519195
Family Sd	0.0597549	0.0541210	0.0510351	0.0487231	0.0464730	0.0475891	0.0464657
Health Sd	0.0842027	0.0788170	0.0695582	0.0666190	0.0627531	0.0671350	0.0656782
Freedom Sd	0.0367217	0.0321067	0.0289789	0.0273835	0.0274698	0.0274818	0.0271782
Generosity Sd	0.0147743	0.0124477	0.0098803	0.0092511	0.0089464	0.0088534	0.0088469
Trust Sd	0.0238065	0.0192423	0.0171973	0.0163574	0.0152388	0.0142809	0.0156814

Random Forest

While an RF is said to be self-validating, given the nature of the algorithm, 1000 70/30 train/test splits were done to estimate the MSE and σ of the estimates. Hyperparameter optimization was conducted on the number of trees in the model The following number of trees were used: 50, 100, 250, 500, 1000, 1500, 2000. The results are as follows:

The optimum RF, the one with the lowest MSE that does not overfit, for this data contains 1000 trees.

Furthermore, variable importance was extracted from the random forest model.

It is clear from Table that Health, Economy, Family and Freedom, in that order play an important role in predicting happiness in a country. Generosity and Trust appear to lack much predictive power. Interpreting these results, the most important factor in predicting happiness is how long and healthy an individual is during their life. After that how wealthy, then how much support, then how free an individual is to make choices, round out the most important predictors. Variable importance was also consistent across different numbers of trees.

If we were to engage in sophomoric psychology, we would find these correspond rather well to Maslow's Hierarchy of needs. [5] The most important need is that of corporeal safety, i.e., safety in one's body. Following that economic safety, the ability to earn enough money to support one's self. Then moving on to the ability to find aid in times of trouble, and at last being free to choose how one uses their life.

[5]Maslow, A (1954). Motivation and personality. New York, NY: Harper.

Table 7: Support Vector Machine Results

	1e-04	0.005	0.001	0.01	0.1	0.5	1
Mean MSE Train	0.2384972	0.2393401	0.2375916	0.2391010	0.2372099	0.2381784	0.2375120
Mean MSE Test	0.2694016	0.2675648	0.2711220	0.2671151	0.2723666	0.2692470	0.2712271
Mean MSE Sd Train	0.0224398	0.0229875	0.0224430	0.0223839	0.0220509	0.0231101	0.0225539
Mean MSE Sd Test	0.0551869	0.0551480	0.0542401	0.0540470	0.0537209	0.0545778	0.0557396

Support Vector Machine

Two SVMs were fit to the data. The first was an SVM with radial kernel. This model was found to overfit the data and so was replaced with a model with linear kernel. This model did not overfit the data. 1000 70/30 train/test splits were done to estimate the MSE and the σ of the estimates. Hyperparameter optimization was conduced on ϵ, the insensitive-loss function. The following values were used for ϵ: 10^{-4}, 0.005, 0.001, 0.01, 0.1, 0.5, 1

From the above table, the optimum ϵ is then .005.

Table 8: Best Models

	LASSO	RF: 1000	SVM: .005
MSE	0.2721838	0.2777539	0.2675648
MSE Sd	0.0545674	0.0514747	0.0551480

Optimum Model

The results for the three different algorithms show that, first, none is significantly better than the rest at predicting national happiness, the best MSEs for each algorithm being close in value. However, the lowest value came from the LASSO. This model possessed the lowest MSE for the test data.

Conclusions

Three main conclusions can be drawn the analysis performed in this report. These conclusions are: Health, Economy, and Family are the most important predictors of a nation's happiness, there exist a non-zero number of nations whose happiness does not follow the patterns seen in other nations, and the nations of the world can be divided into three groups in terms of the predictors identified in this report. These conclusions will be discussed.

As can be seen from the RF results, three predictors are far more important in modeling happiness than the others. Those three predictors, Health, Economy and Family, are disproportionately able to predict a nation's happiness. The reasons for this are many fold, though they can be explained in terms of basic human needs: happiness is impossible without health, economic stability, and the ability to recover from adversity. For these reasons it is not too surprising that these predictors have such greater importance in the model.

There exist certain nations, identified by the bi-plot, whose happiness cannot be adequately explained in the same manner as other nations. Some of these nations, namely Somalia and Taiwan, need further study to determine why their happiness is where it is. In the case of Somalia further study is needed on the validity and accuracy of economic data from the country. For Taiwan further study is needed to determine why the country's happiness is so high given its high level of corruption.

Lastly, the PCA, the K-Means Clustering, and the Bi-Plot, all showed that the nations of the world can be divided into 3 groups: the happy, the unhappy, and the ambivalent. This distinction conforms well to the developed, developing and underdeveloped tripartite in geopolitical theory.