Syracuse University

World Happiness Data Analysis

Sandra Tang, Nicole Villanueva, Kate Walko, Emma Woods

2019-1002 IST 707-17401 Data Analytics

Dr. Ami Gates

December 12th, 2019

TABLE OF CONTENTS

INTRODUCTION

The definition of happiness varies widely from person to person and from culture to culture. At the individual level, it is considered a positive personal experience based on emotional (positive and negative affect) and cognitive (mental health) interpretations of life events and qualities. At the collective level, the happiness of a given population or region or country can be calculated as the sum and/or averages of the reflections and assessments of the individuals. The importance of measuring happiness on a greater scale has risen in recent years as many studies have found that traditional evaluations of economic growth, education level, and the distribution of wealth in a given population are not sufficient indicators for quality of life on their own.

Although the world is wealthier on average, unhappiness is remarkably prevalent, even in the wealthiest countries. The heavily unbalanced distribution of wealth actually accounts for much of the discrepancies in happiness, and there is strong evidence that increasing an individual's wealth does not proportionately increase personal happiness. Personal wealth increases happiness to the extent that it can secure basic needs and comforts; but, increasing wealth beyond that threshold does not equate to an equal increase in personal happiness. Conversely, the obsession over wealth and economic growth can actually create unhappiness. The World Happiness Report 2019 asserts that this unhappiness can manifest as detrimental behaviors like substance abuse, addiction, risky behavior such as gambling, and uninhibited spending. Therefore, happiness must be measured by many factors beyond wealth, such as the length of healthy life expectancy, trust in the government, and freedom to make life choices.

By analyzing the happiness of a country through the average or sum of individual assessments of factors including social support, corruption perception, and generosity, more and more countries are able to understand how adjustments to public policy can affect the quality of life within their given country or population. Generally, the goal is to promote increasing levels of happiness in order to ensure healthier, longer lives, and satisfaction with living conditions.

Measuring happiness across entire countries is a new development that can now be rigorously evaluated. Through a multitude of methods such as self-report surveys, behavior evaluations, brain imaging, and advancements in machine learning, people in charge of informing public policy can now understand where a population may have room for improvement and/or identify areas where they can affect the most positive change. It has become abundantly clear that, while individual happiness is certainly one's own top priority, the collective happiness of a country is critical to the success and well-being of citizens, a country's government, private and public companies and schools, and beyond.

ANALYSIS

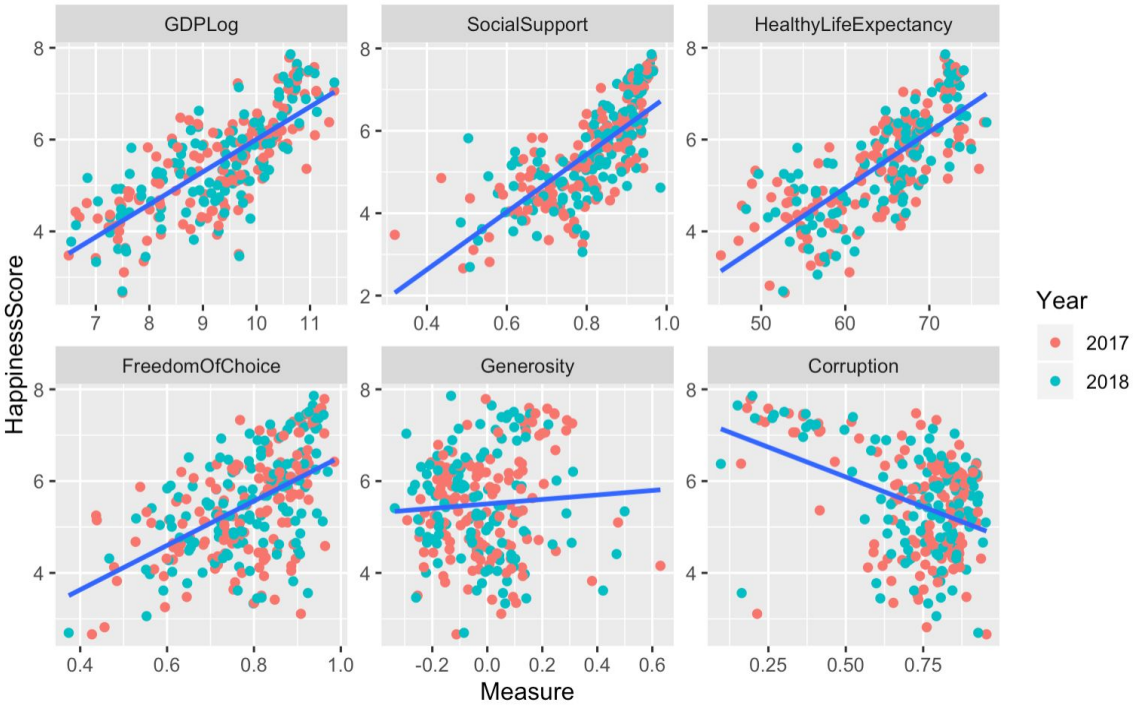
About the Data

The dataset is obtained from the 2019 World Happiness Report which contains empirical linkages between a number of national measures of the quality of government and national average happiness. Below are descriptions about the principal measures of happiness (life evaluations, positive affect, and negative affect) along with their main supporting factors, which are GDP per capita, social support, healthy life expectancy, freedom, generosity, and absence of corruption. Happiness measures in each country are not constructing using these six factors but are based on individuals' own assessments of their lives, as indicated by the Cantril ladder. Therefore, the six variables measured in the report can help explain the variation of happiness across countries. The 2019 World Happiness Report and dataset can be found here: https://worldhappiness.report/ed/2019.

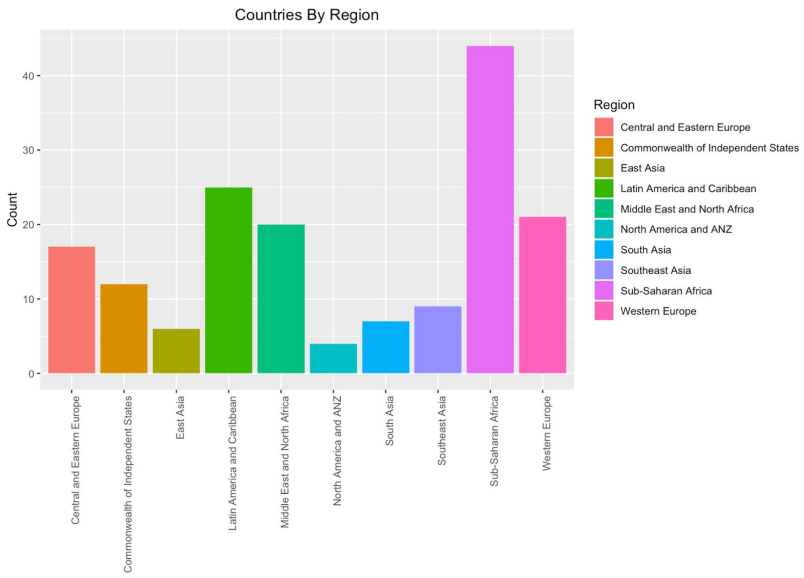
Data Element	Definition	Data Source
Country name	Name of 165 countries	
Year	Year of the data ranging from 2005 to 2018	
Life Ladder	Happiness score or subjective well-being. Life evaluation used is the Cantril Ladder, which asks survey respondents to place the status of their lives on a "ladder" scale ranging from 0 to 10, where 0 means the worst possible life and 10 means the best possible life.	Gallup World Poll (GWP)
Positive affect	Average of previous-day affect measures for happiness, laughter, and enjoyment for GWP waves 3-7 (years 2008 to 2012, and some in 2013). It is defined as the average of laughter and enjoyment for other waves where the happiness question was not asked. The general form for the affect questions is: Did you experience the following feelings during a lot of the day yesterday?	Gallup World Poll (GWP)
Negative affect	The average of previous-day affect measures for worry, sadness, and anger for all waves.	Gallup World Poll (GWP)
Log GDP per capita	Natural log of GDP per capita in purchasing power parity (PPP) at constant 2011 international dollar prices	World Development Indicators (WDI)
Social support	National average of the binary responses (either 0 or 1) to GWP 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?"	Gallup World Poll (GWP)

Healthy life expectancy at birth	Healthy life expectancy at birth constructed based on data available for 2005, 2010, 2015, and 2016. To match the report's sample period, interpolation and extrapolation are used.	World Health Organization's (WHO) Global Health Observatory data repository
Freedom to make life choices	National average of responses to GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"	Gallup World Poll (GWP)
Generosity	Residual of regressing national average of response to GWP question "Have you donated money to a charity in the past month?" on GDP per capita	Gallup World Poll (GWP)
Perceptions of corruption	National average of survey responses to two questions in the GWP: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?"	Gallup World Poll (GWP)

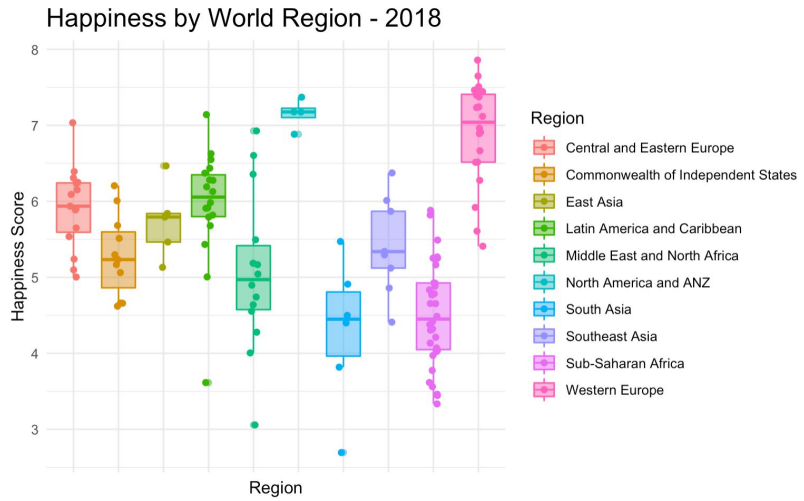
The following correlation plots below illustrate how each of the six variables (x axis) measured in the report relate to the overall happiness score (y-axis) for all countries in 2017 (red) and 2018 (blue). GDP, social support, healthy life expectancy, freedom of choices, and generosity all have positive correlation to happiness score. The relationship between perceptions of corruption and happiness is inverse because higher level of happiness is driven by lower perceived corruption.



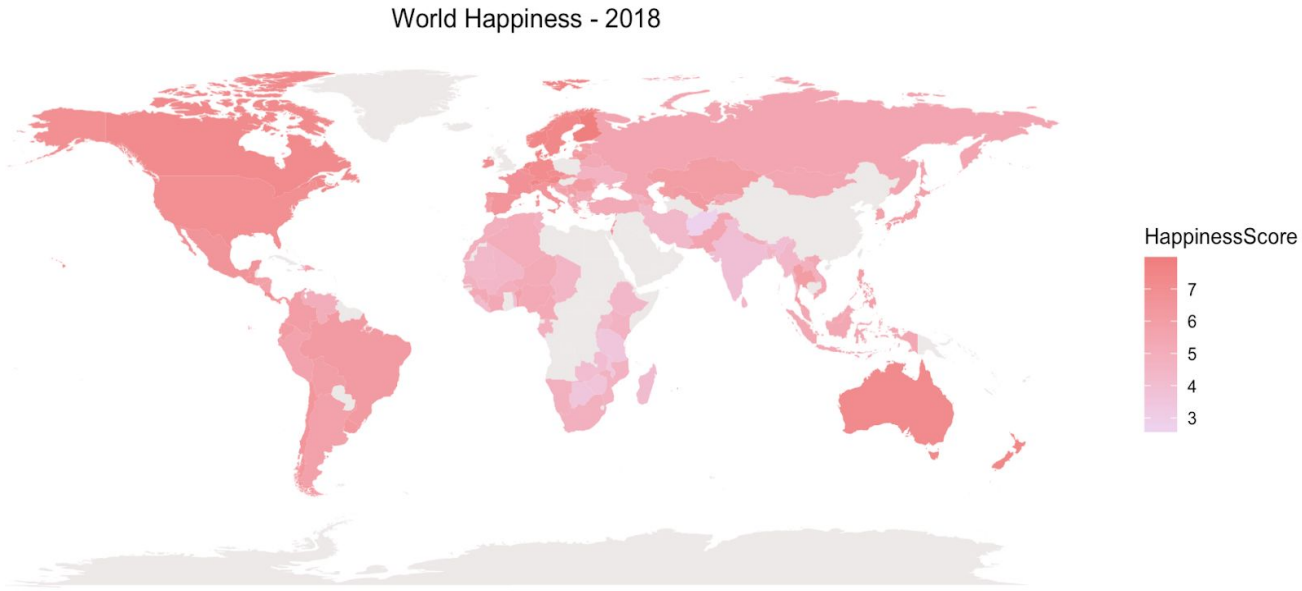
To supplement the world happiness report data, each country is also identified by its corresponding world region obtained from past World Happiness Reports. Viewing happiness by world regions enables a better understanding of happiness at a broader geographical level. Below is a histogram showing the world regions used by the World Happiness Report and the number of countries represented in each region:



Further exploring regionality of countries and happiness reveals an interesting pattern in happiness scores. The 2019 World Happiness dataset shows that, overall, there are many happy countries in Western Europe (Finland, Denmark, Switzerland, Netherlands, Norway), followed by the region grouping of Australia, New Zealand, and North America. Certain South Asia (Afghanistan and India) and several Sub-Saharan Africa countries (Malawi, Tanzania, Botswana, Rwanda) have the lowest happiness scores and are considered the least happy. The boxplot below shows 2018 happiness scores' distribution by world regions.



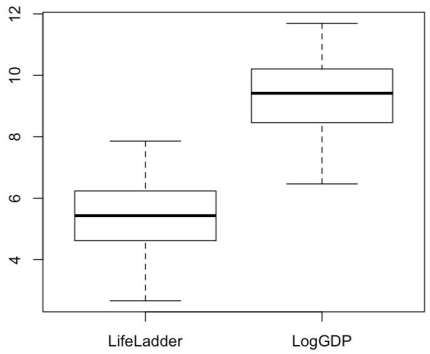
The 2018 happiness scores for each country can also be visualized using a world map to easily identify countries and regions of varying levels of happiness. Areas in grey in the map below correspond to countries that do not have happiness scores for year 2018 in the World Happiness Report.

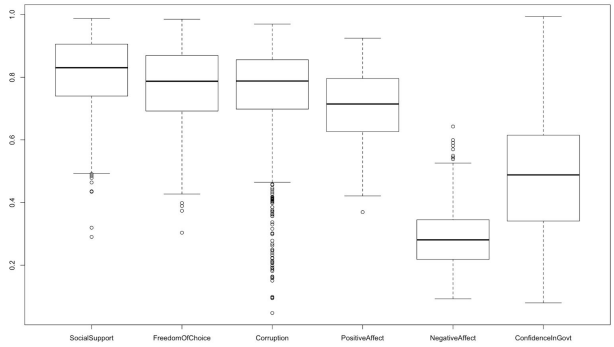


Association Rule Mining Model

To prepare the data for association rule mining, the numerical values for the 10 most influential variables on happiness needed to be discretized and converted into nominal variables. The 10 most influential variables are: life ladder, log of GDP per capita, social support, healthy life expectancy, freedom of choice, generosity, corruption, positive affect, negative affect, and confidence in government. To discretize the variables, it was critical to first outline the distribution of the variables by column:

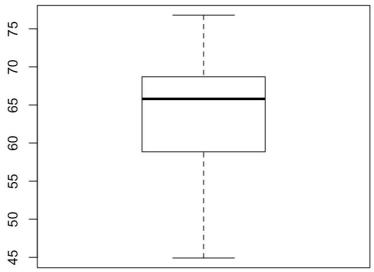
Variables such as life ladder and the log of GDP were distributed as values from 0-12. The intervals were based on logical approximations around the first and third quartile. Values equal or less than the first quartile were considered “low,” values equal or less than the third quartile were considered “moderate” or “medium,” and values greater than the third quartile were considered “high.”



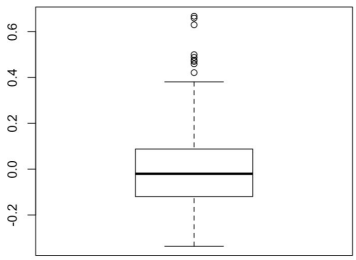


Social support, freedom of choice, corruption, positive/negative affect, and confidence in government were averages of responses to binary responses so the values ranged from 0 to 1. The intervals were determined by the first quartile and the mean or median and marked with “low,” “moderate”/“medium,” or “high.” The intervals for confidence in government were determined by the first and third quartiles.

The value range for healthy life expectancy was from about 45 to 75, and the intervals were determined by the first quartile and the median. The spread of values below the median was greater than the spread above the median.



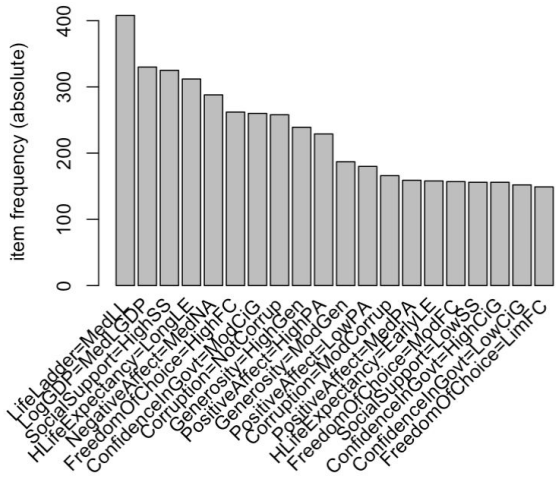
Healthy Life Expectancy



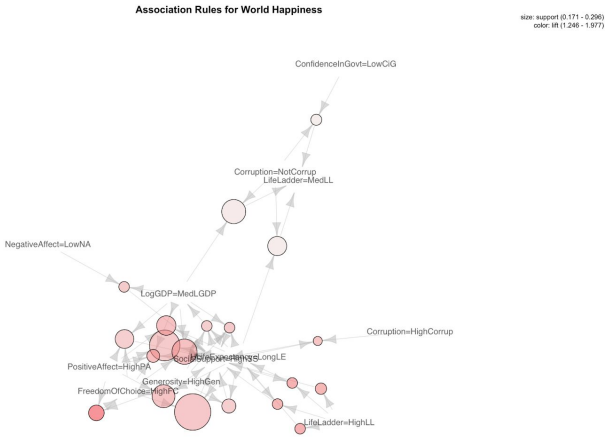
Generosity

For generosity, the values were an average of the binary responses between two questions, so the range was between -0.5 and 0.7. The intervals were determined by the first quartile and the mean.

Once the data was nominal rather than numerical, it was converted into factors and then converted into transactions. It was then useful to observe the item frequency plot to identify trends in the data. Of the 568 observations, 400 had a medium estimation of the life ladder and a majority of observations had medium measurements of log of GDP per capita as well as a high sense of social support.



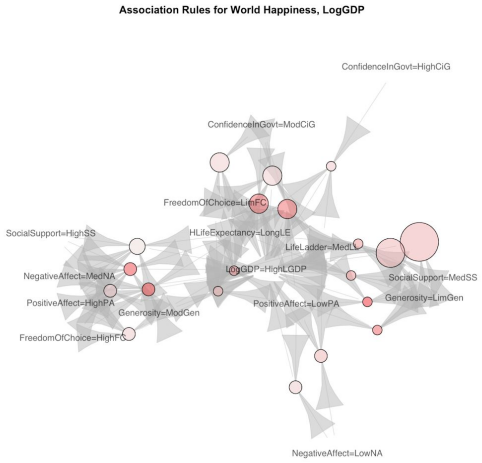
The base association rule model ran without a predetermined rhs:



Key rules to note:

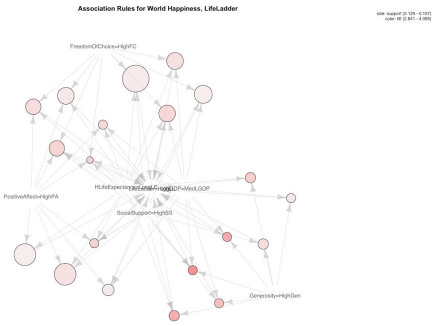
	lhs		rhs	support	confidence	lift	count
[1]	{HLifeExpectancy=LongLE,FreedomOfChoice=HighFC}	=>	{SocialSupport=HighSS}	0.2957746	0.9385475	1.640292	168
[2]	{HLifeExpectancy=LongLE,PositiveAffect=HighPA}	=>	{SocialSupport=HighSS}	0.2693662	0.9272727	1.620587	153
[3]	{LogGDP=MedLGDP,HLifeExpectancy=LongLE,FreedomOfChoice=HighFC}	=>	{SocialSupport=HighSS}	0.2447183	0.9586207	1.675374	139
[4]	{LogGDP=MedLGDP,Corruption=NotCorrup}	=>	{LifeLadder=MedLL}	0.2394366	0.9189189	1.279279	136
[5]	{HLifeExpectancy=LongLE,FreedomOfChoice=HighFC,PositiveAffect=HighPA}	=>	{SocialSupport=HighSS}	0.2341549	0.9500000	1.660308	133
[6]	{LogGDP=MedLGDP,HLifeExpectancy=LongLE,PositiveAffect=HighPA}	=>	{SocialSupport=HighSS}	0.2183099	0.9612403	1.679952	124
[7]	{HLifeExpectancy=LongLE,Corruption=NotCorrup}	=>	{LifeLadder=MedLL}	0.2165493	0.9044118	1.259083	123
[8]	{LogGDP=MedLGDP,FreedomOfChoice=HighFC,PositiveAffect=HighPA}	=>	{SocialSupport=HighSS}	0.2147887	0.8970588	1.567783	122
[9]	{Generosity=HighGen,PositiveAffect=HighPA}	=>	{FreedomOfChoice=HighFC}	0.2007042	0.9120000	1.977160	114
[10]	{HLifeExpectancy=LongLE,Generosity=HighGen}	=>	{SocialSupport=HighSS}	0.1936620	0.9016393	1.575788	110
[11]	{LogGDP=MedLGDP,HLifeExpectancy=LongLE,FreedomOfChoice=HighFC,PositiveAffect=HighPA}	=>	{SocialSupport=HighSS}	0.1883803	0.9727273	1.700028	107
[12]	{LifeLadder=HighLL}	=>	{SocialSupport=HighSS}	0.1813380	1.0000000	1.747692	103
[13]	{Corruption=NotCorrup,ConfidenceInGovt=LowCiG}	=>	{LifeLadder=MedLL}	0.1795775	0.8947368	1.245614	102
[14]	{LifeLadder=HighLL,HLifeExpectancy=LongLE}	=>	{SocialSupport=HighSS}	0.1778169	1.0000000	1.747692	101
[15]	{LifeLadder=HighLL}	=>	{HLifeExpectancy=LongLE}	0.1778169	0.9805825	1.785163	101
[16]	{LifeLadder=HighLL,SocialSupport=HighSS}	=>	{HLifeExpectancy=LongLE}	0.1778169	0.9805825	1.785163	101
[17]	{LogGDP=MedLGDP,HLifeExpectancy=LongLE,Generosity=HighGen}	=>	{SocialSupport=HighSS}	0.1778169	0.9351852	1.634416	101
[18]	{LogGDP=MedLGDP,NegativeAffect=LowNA}	=>	{SocialSupport=HighSS}	0.1778169	0.9266055	1.619421	101
[19]	{SocialSupport=HighSS,HLifeExpectancy=LongLE,Generosity=HighGen}	=>	{LogGDP=MedLGDP}	0.1778169	0.9181818	1.580386	101
[20]	{HLifeExpectancy=LongLE,Corruption=HighCorrup}	=>	{SocialSupport=HighSS}	0.1707746	0.9509804	1.662021	97

The second association rule model had HighLGDP set for the rhs:



	lhs		rhs	support	confidence	lift	count
[1]	{SocialSupport=MedSS,Generosity=LimGen}	=>	{LogGDP=HighLGDP}	0.03345070	0.5588235	3.488041	19
[2]	{LifeLadder=MedLL,SocialSupport=MedSS,Generosity=LimGen}	=>	{LogGDP=HighLGDP}	0.02816901	0.5517241	3.443729	16
[3]	{HLifeExpectancy=LongLE,FreedomOfChoice=LimFC,PositiveAffect=LowPA,ConfidenceInGovt=ModCiG}	=>	{LogGDP=HighLGDP}	0.02288732	0.6190476	3.863946	13
[4]	{LifeLadder=MedLL,HLifeExpectancy=LongLE,FreedomOfChoice=LimFC,PositiveAffect=LowPA,ConfidenceInGovt=ModCiG}	=>	{LogGDP=HighLGDP}	0.02288732	0.6190476	3.863946	13
[5]	{HLifeExpectancy=LongLE,FreedomOfChoice=LimFC,ConfidenceInGovt=ModCiG}	=>	{LogGDP=HighLGDP}	0.02288732	0.5200000	3.245714	13
[6]	{LifeLadder=MedLL,HLifeExpectancy=LongLE,FreedomOfChoice=LimFC,ConfidenceInGovt=ModCiG}	=>	{LogGDP=HighLGDP}	0.02288732	0.5200000	3.245714	13
[7]	{SocialSupport=HighSS,HLifeExpectancy=LongLE,Generosity=ModGen,PositiveAffect=HighPA,NegativeAffect=MedNA}	=>	{LogGDP=HighLGDP}	0.02112676	0.5000000	3.120879	12
[8]	{HLifeExpectancy=LongLE,FreedomOfChoice=HighFC,Generosity=ModGen,PositiveAffect=HighPA,NegativeAffect=MedNA}	=>	{LogGDP=HighLGDP}	0.01936620	0.6470588	4.038785	11
[9]	{SocialSupport=HighSS,HLifeExpectancy=LongLE,FreedomOfChoice=HighFC,Generosity=ModGen,PositiveAffect=HighPA,NegativeAffect=MedNA}	=>	{LogGDP=HighLGDP}	0.01936620	0.6470588	4.038785	11
[10]	{LifeLadder=MedLL,PositiveAffect=LowPA,NegativeAffect=LowNA}	=>	{LogGDP=HighLGDP}	0.01936620	0.5500000	3.432967	11
[11]	{PositiveAffect=LowPA,NegativeAffect=LowNA}	=>	{LogGDP=HighLGDP}	0.01936620	0.5238095	3.269492	11
[12]	{FreedomOfChoice=HighFC,Generosity=ModGen,PositiveAffect=HighPA,NegativeAffect=MedNA}	=>	{LogGDP=HighLGDP}	0.01936620	0.5238095	3.269492	11
[13]	{SocialSupport=HighSS,FreedomOfChoice=HighFC,Generosity=ModGen,PositiveAffect=HighPA,NegativeAffect=MedNA}	=>	{LogGDP=HighLGDP}	0.01936620	0.5238095	3.269492	11
[14]	{LifeLadder=MedLL,SocialSupport=MedSS,Generosity=LimGen,PositiveAffect=LowPA}	=>	{LogGDP=HighLGDP}	0.01760563	0.6666667	4.161172	10
[15]	{SocialSupport=MedSS,Generosity=LimGen,PositiveAffect=LowPA}	=>	{LogGDP=HighLGDP}	0.01760563	0.6250000	3.901099	10
[16]	{SocialSupport=MedSS,HLifeExpectancy=LongLE,Generosity=LimGen}	=>	{LogGDP=HighLGDP}	0.01760563	0.5882353	3.671622	10
[17]	{LifeLadder=MedLL,SocialSupport=MedSS,HLifeExpectancy=LongLE,Generosity=LimGen}	=>	{LogGDP=HighLGDP}	0.01760563	0.5882353	3.671622	10
[18]	{LifeLadder=MedLL,HLifeExpectancy=LongLE,FreedomOfChoice=LimFC,Generosity=ModGen,PositiveAffect=LowPA}	=>	{LogGDP=HighLGDP}	0.01760563	0.5882353	3.671622	10
[19]	{HLifeExpectancy=LongLE,FreedomOfChoice=LimFC,Generosity=ModGen,PositiveAffect=LowPA}	=>	{LogGDP=HighLGDP}	0.01760563	0.5555556	3.467643	10
[20]	{LifeLadder=MedLL,HLifeExpectancy=LongLE,ConfidenceInGovt=HighCiG}	=>	{LogGDP=HighLGDP}	0.01760563	0.5263158	3.285136	10

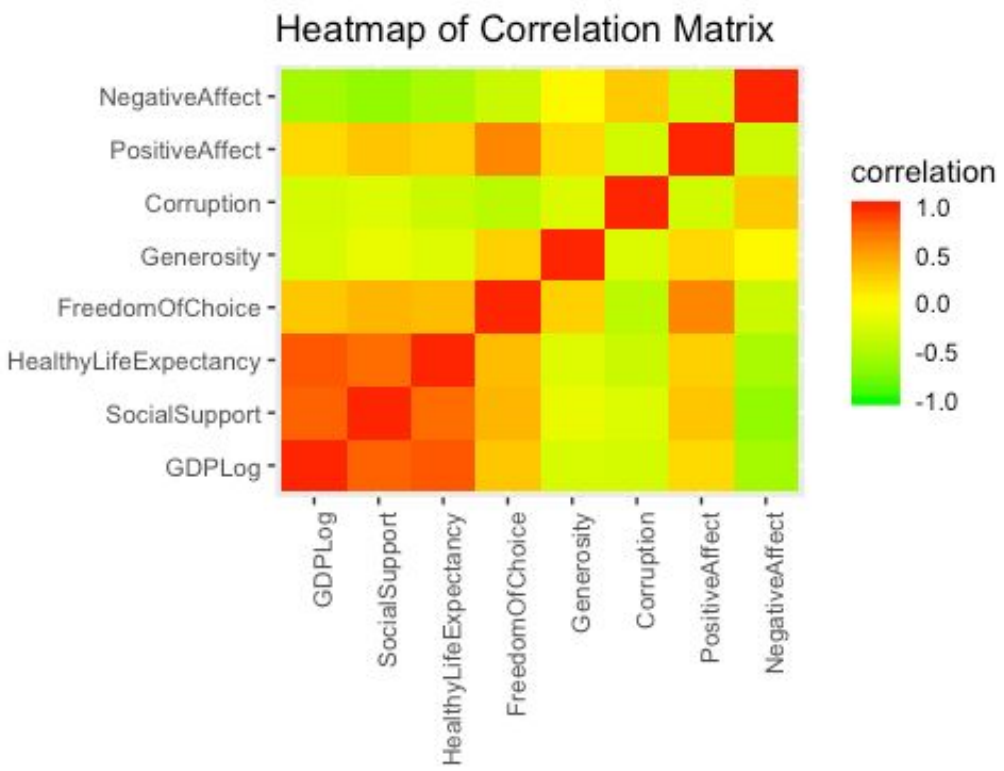
The third association rule model had HighLL (high life ladder) set for the rhs:



lhs		rhs	support	confidence	lift	count
{SocialSupport=HighSS,HLifeExpectancy=LongLE,FreedomOfChoice=HighFC}	=>	{LifeLadder=HighLL}	0.1566901	0.5297619	2.921405	89
{SocialSupport=HighSS,HLifeExpectancy=LongLE,PositiveAffect=HighPA}	=>	{LifeLadder=HighLL}	0.1496479	0.5555556	3.063646	85
{HLifeExpectancy=LongLE,PositiveAffect=HighPA}	=>	{LifeLadder=HighLL}	0.1496479	0.5151515	2.840836	85
{LogGDP=MedLGDP,SocialSupport=HighSS,FreedomOfChoice=HighFC}	=>	{LifeLadder=HighLL}	0.1443662	0.5157233	2.843989	82
{LogGDP=MedLGDP,SocialSupport=HighSS,HLifeExpectancy=LongLE,FreedomOfChoice=HighFC}	=>	{LifeLadder=HighLL}	0.1426056	0.5827338	3.213522	81
{LogGDP=MedLGDP,HLifeExpectancy=LongLE,FreedomOfChoice=HighFC}	=>	{LifeLadder=HighLL}	0.1426056	0.5586207	3.080549	81
{SocialSupport=HighSS,FreedomOfChoice=HighFC,PositiveAffect=HighPA}	=>	{LifeLadder=HighLL}	0.1426056	0.5328947	2.938682	81
{SocialSupport=HighSS,HLifeExpectancy=LongLE,FreedomOfChoice=HighFC,PositiveAffect=HighPA}	=>	{LifeLadder=HighLL}	0.1408451	0.6015038	3.317030	80
{HLifeExpectancy=LongLE,FreedomOfChoice=HighFC,PositiveAffect=HighPA}	=>	{LifeLadder=HighLL}	0.1408451	0.5714286	3.151179	80
{LogGDP=MedLGDP,SocialSupport=HighSS,PositiveAffect=HighPA}	=>	{LifeLadder=HighLL}	0.1355634	0.5238095	2.888581	77
{SocialSupport=HighSS,HLifeExpectancy=LongLE,Generosity=HighGen}	=>	{LifeLadder=HighLL}	0.1338028	0.6909091	3.810062	76
{HLifeExpectancy=LongLE,Generosity=HighGen}	=>	{LifeLadder=HighLL}	0.1338028	0.6229508	3.435302	76
{SocialSupport=HighSS,Generosity=HighGen}	=>	{LifeLadder=HighLL}	0.1338028	0.5588235	3.081668	76
{LogGDP=MedLGDP,SocialSupport=HighSS,HLifeExpectancy=LongLE,Generosity=HighGen}	=>	{LifeLadder=HighLL}	0.1320423	0.7425743	4.094973	75
{LogGDP=MedLGDP,HLifeExpectancy=LongLE,Generosity=HighGen}	=>	{LifeLadder=HighLL}	0.1320423	0.6944444	3.829558	75
{LogGDP=MedLGDP,SocialSupport=HighSS,Generosity=HighGen}	=>	{LifeLadder=HighLL}	0.1320423	0.6521739	3.596454	75
{LogGDP=MedLGDP,SocialSupport=HighSS,HLifeExpectancy=LongLE,PositiveAffect=HighPA}	=>	{LifeLadder=HighLL}	0.1320423	0.6048387	3.335421	75
{LogGDP=MedLGDP,HLifeExpectancy=LongLE,PositiveAffect=HighPA}	=>	{LifeLadder=HighLL}	0.1320423	0.5813953	3.206141	75
{LogGDP=MedLGDP,Generosity=HighGen}	=>	{LifeLadder=HighLL}	0.1320423	0.5357143	2.954230	75
{LogGDP=MedLGDP,SocialSupport=HighSS,FreedomOfChoice=HighFC,PositiveAffect=HighPA}	=>	{LifeLadder=HighLL}	0.1285211	0.5983607	3.299698	73

Clustering Model

To identify how countries group into ‘happy’ and ‘unhappy’ segments (and variations in-between), the dataset is subsetted, in order to extrapolate 2018 data only. After scaling the main predictor variables, a correlation matrix may be visualized using a heatmap in order to identify relationships between the variables to be clustered:

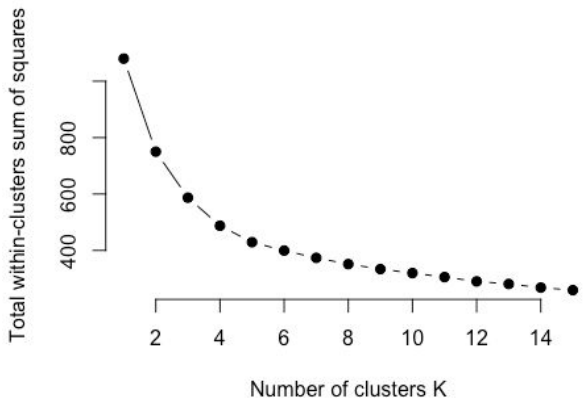


A heatmap correlation matrix showing the degree of correlation with happiness scores versus each component measured in the 2018 World Happiness Report.

GDP (“GDPLog”) and social factors such as Healthy Life Expectancy and Social Support, appear to have a high degree of positive correlation.

K-means Clustering

Before applying the k-means algorithm, an effective number of means (clusters) was assessed. An elbow plot was generated to identify the appropriate number of clusters with which clustering analysis could then be conducted:

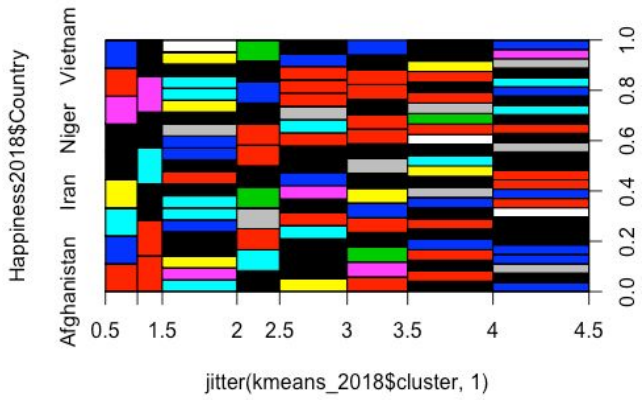


This plot reveals that the number of clusters for K drops precipitously after two clusters and continues to drop.

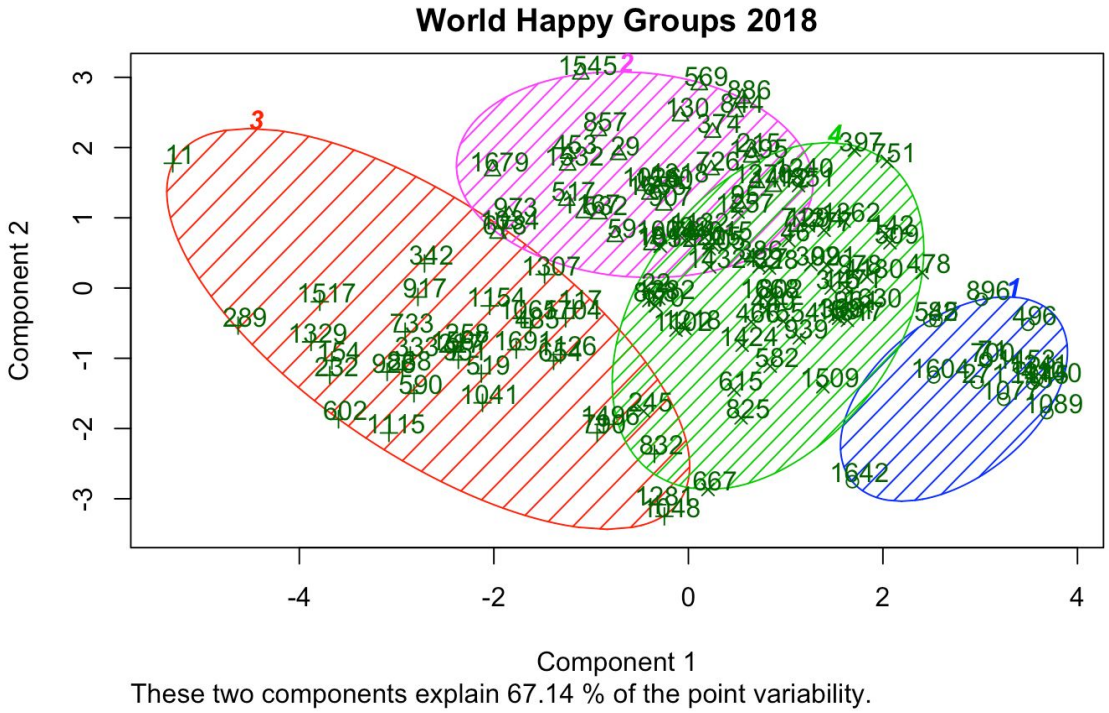
The elbow plot indicates a fairly significant drop up until 4 clusters, where it then starts to slowly level off. It was identified that 4 clusters should provide meaningful results for analysis. The kmeans algorithm was then applied using k = 4 to select 4 cluster groups. Only the main predictor variables were included for k-means cluster analysis, these being the following;

- GDP Log
- Social Support
- Healthy Life Expectancy
- Freedom of Choice
- Generosity
- Corruption
- Positive Affect
- Negative Affect

Happiness Scores in 2018 by Country analyzed by k-means algorithm clustering.



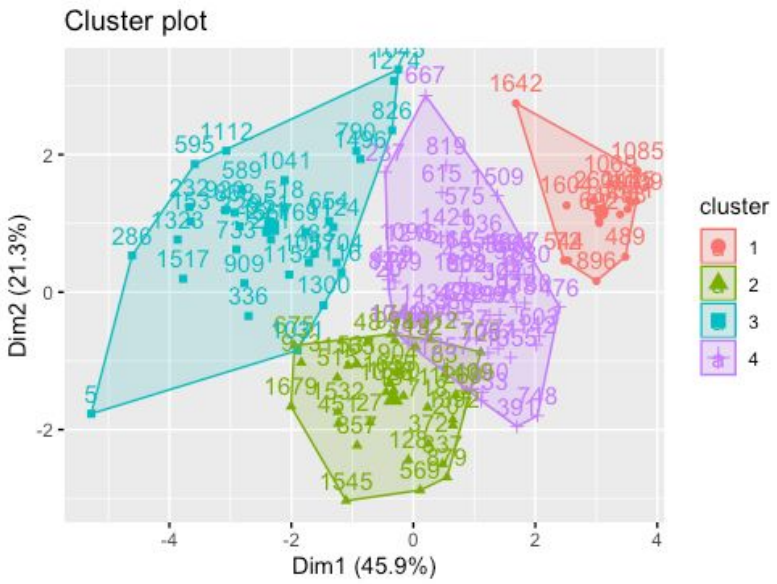
After applying Principal Component Analysis (PCA) in order to reduce dimensionality in the data, we see that 4 fairly distinct clusters are apparent as shown in the figure below:



These two components explain 67.14 % of the point variability.

PCA helps identify four different clusters in the Happiness Score data for 2018's scores.

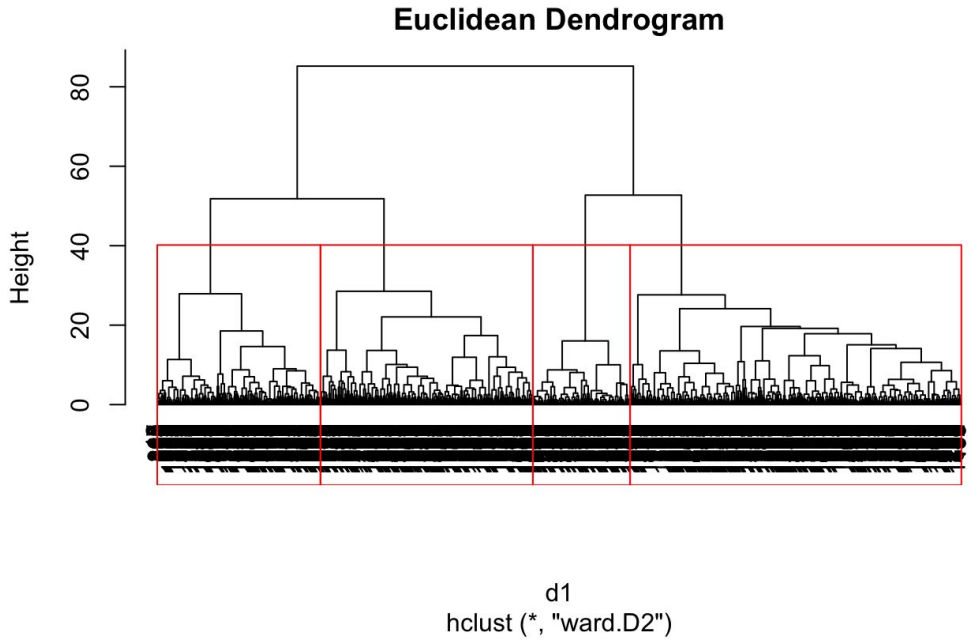
Further cluster plotting clearly shows that the k-means algorithm successfully identified four distinct segments with virtually no overlap, shown in figure 8 below:



Plotting k-means clustering results reveals the algorithm grouped scores into four distinct clusters.

Hierarchical Clustering

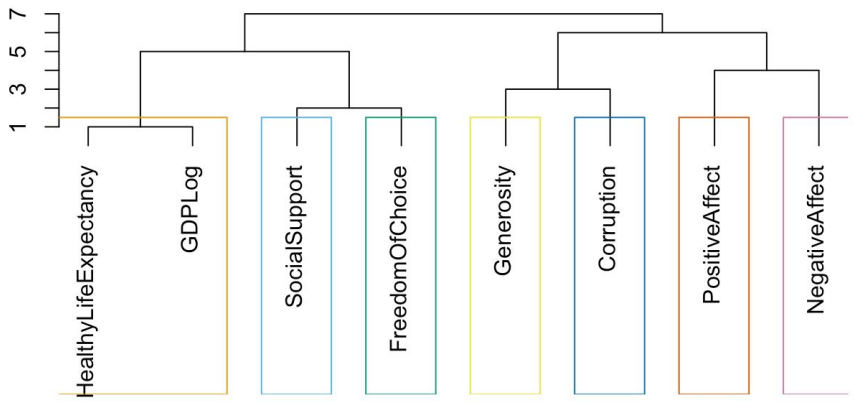
Before applying hierarchical clustering techniques to the data, a Euclidean distance matrix was first created. Applying the hclust algorithm using the distance matrix results and utilizing the “ward.D2” results in the following dendrogram shown in the following dendrogram.



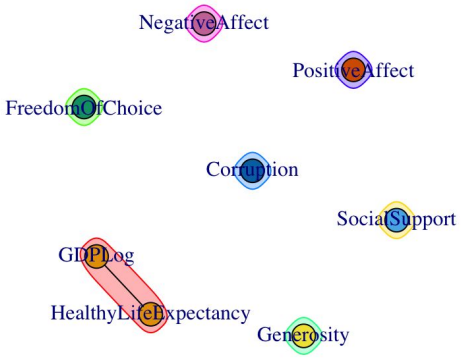
A dendrogram created using hierarchical clustering; four distinct clusters emerge in the resulting dendrogram, the same as clustering revealed with the k-means algorithm.

A Cosine Similarity Matrix is then created to facilitate comparison of the application of Cosine Similarity as a distance measure with the results from Euclidean distance hierarchical clustering. In order to apply Cosine Similarity measures, the data is converted to matrix format. Cosine Similarity is then applied with the edges of the tree pruned at 0.70 to ensure results would be contained and easier to interpret. Using this method, we can see the world happiness predictor factors clustered as follows:

Plotting the cosine similarity as a distance measure.



It is interesting to note that factors “Healthy Life Expectancy” and “GDP Log” appear clustered together in one group, indicating there is a strong correlation between the two variables. Generating a graph using an adjacency matrix and calculating the edge betweenness of the network further provides the ability to plot a network model of these groups and assess the community structure within the network. Such a graph reiterates the observation that there is a strong link between GDP Log and Healthy Life Expectancy, as shown below:



A network model plot of each of the World Happiness Report data (all years).

The color coding scheme from the dendrogram shown below enables us to see that Positive and Negative affect are correlated, but that positive affect is also related to social support. Social support also appears to have ties to Generosity.

Decision Tree Model

As mentioned in the “About the Data” section, the happiness score of a country is correlated with the supporting factors of GDP, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption. Therefore, these factors can help predict the happiness of a country up to a certain degree. Decision tree is a classification algorithm that can be used to create a prediction model with the happiness score as a target response while the 6 supporting factors serve as predictor variables. The resulting model can then be leveraged to predict the happiness of a country, given different values GDP, social support, healthy life expectancy, freedom of life choices, generosity, and perceptions of corruption.

Decision trees are capable of performing both regression and classification tasks. For the purpose of this analysis, the decision tree model will be fitted as a classification one. This means the happiness score, which is of numeric type, is first discretized into 3 classes of happiness levels (high, medium, and low) to be used as labels of type factor. In order to predict the future happiness level of a country, the training data are comprised of complete historical happiness data from years 2013 to 2017, while data from 2018 is for testing and evaluation of the model accuracy.

Prior to fitting a decision tree model, features evaluation is first executed using 3 different methods: gain ratio, gini-index, and information gain. The output below informs the relevance and importance of the 6 predictor attributes by each splitting measure. All 3 evaluation functions appear similar, listing healthy life expectancy, GDP, and social support as having relevance than the other 3 features.

	Gini <dbl>
HealthyLifeExpectancy	0.14015033
GDPLog	0.12729688
SocialSupport	0.11774030
Corruption	0.07881161
FreedomOfChoice	0.07326811
Generosity	0.02836685

	InfoGain <dbl>
HealthyLifeExpectancy	0.41887199
GDPLog	0.33499080
SocialSupport	0.32115668
FreedomOfChoice	0.16585846
Corruption	0.16319877
Generosity	0.05953287

	GainRatio <dbl>
HealthyLifeExpectancy	0.4647008
GDPLog	0.4043333
SocialSupport	0.3369761
Corruption	0.2755587
FreedomOfChoice	0.2432301

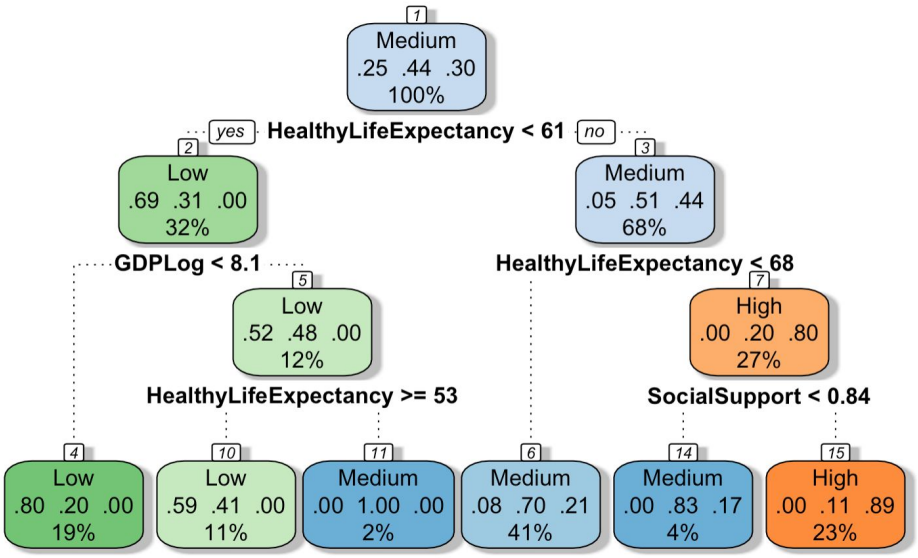
The first iteration of the decision tree accuracy is about 69.5%. Replacing the default splitting function of “gini” with “Information Gain” achieves a slight accuracy improvement to 72%. To obtain a simpler tree, the model is also instructed with the argument that any split which does not improve the fit by cp of 0.01 be pruned off by cross-validation and not to be pursued. The decision tree summary indicates that the highest variables of influence are healthy life expectancy, GDP, and social support.

```
Call:
rpart(formula = HappinessLevel ~ ., data = Happiness_prior[,
    !names(Happiness_prior) %in% c("Country", "Year", "PositiveAffect",
        "NegativeAffect", "HappinessScore", "Region", "RangedHLE",
        "SDHLE")], method = "class", parms = list(split = "information"),
    cp = 0.01)
  n= 643

          CP nsplit rel error    xerror       xstd
1 0.25279330      0 1.0000000 1.0000000 0.03518645
2 0.04189944      2 0.4944134 0.5363128 0.03241532
3 0.01396648      3 0.4525140 0.5251397 0.03221782
4 0.01000000      5 0.4245810 0.4776536 0.03129540

Variable importance
HealthyLifeExpectancy               GDPLog        SocialSupport       FreedomOfChoice            Corruption
                   38                   27                   19                    6                    5
           Generosity
                    4
```

Following is the decision tree graph that represents the model choices and results in the form of a tree. The nodes in the graph are the classification outcomes and the edges of the graph represent the decision rules or conditions. The attribute that best classifies the training data is “healthy life expectancy < 61” and can be seen as the root of the tree.



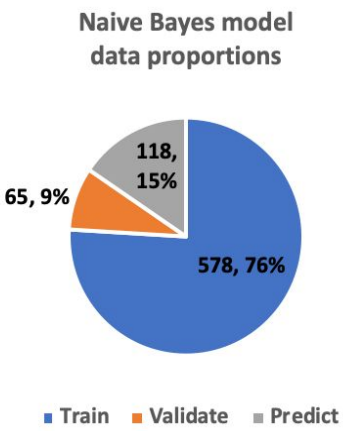
Naive Bayes Model

In order to run the Naive Bayes method for prediction, a similar preprocessing approach to the one applied to the Decision Tree model was used. As a method of supervised learning requiring a class label from which to predict, the numeric values of the happiness scores were discretized into separate categories for use as nominal labels. The happiness score itself for each observation within the data set was not included within the train or test data. The prepared data comprised all complete cases for the happiness data for the years 2013 to 2017, with 2018 observations reserved for use as the prediction ("hold out") set.

It was decided that two different models would be applied, the first being a simple categorization between "happy" and "unhappy" countries. In order to run this version of the model, the Happiness Score was discretized into 2 labels. To evaluate the appropriate splits, both the average and maximum values within the data set were calculated. The mean was found to be 5.40996, with the maximum happiness score applied being 7.97. For future reference, it was noted that the observation in the data frame resulting in this score belonged to **Denmark** in **2008**.

For two-category prediction, the train and test labels were split by characterizing all observations in the training data >4 as "unhappy" and everything < 4 as "happy". A second label set was also created using the same cut points used for Decision Tree, whereby "low" (< 4.5), "medium" (4.5 - 6) and "high" (6+) so that these three happiness score categories could also be predicted and in order to provide a prediction comparison. Both sets of labels were saved for later recall. The predictor variables within the data frame were separated, with 2018 data removed as the 'hold out' test set for final prediction scoring. This data amounts to a total of 118 observations. The remaining years' data (2013 - 2017) were split into a training and testing set. This was achieved by removing 90% of the data for training with the remaining 10% to be used for validation testing of the model.

Proportions of the train/test/holdout data are visualized:



Here the validation data provides a testing set from which to check accuracy scores following application of the Naive Bayes algorithm to the training data. The 'predict' or holdout set contains 2018 data with happiness scores removed, to be used for prediction once the model is ready for final testing and application to the data.

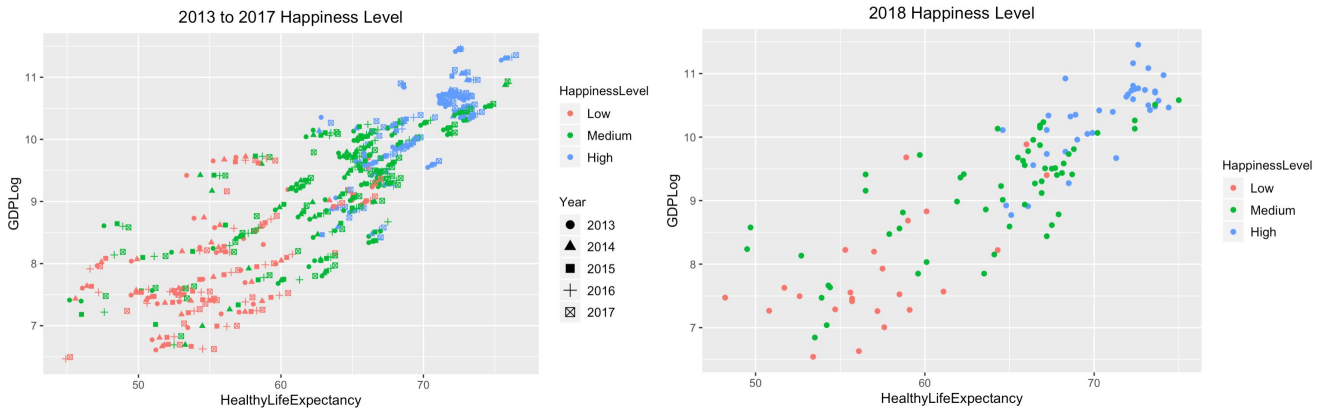
Following initial test results, further experiments are applied in order to see if the accuracy of the model can be improved further. As the Naive Bayes method

assumes that all predictor variables are **independent**, the correlation matrix was analyzed in order to identify those that show a high degree of correlation. As the variables Healthy Life Expectancy and Social Support are highly correlated within the matrix, and it has been observed from exploratory data analysis techniques that Healthy Life Expectancy (HLE) appears to be a strong predictor against happiness score values, the model was run again excluding this variable and instead using the following: GDP Log, Social Support, Freedom of Choice and Corruption. A final version was also run, this time to include HLE and other social support factors (Social Support, Freedom of Choice, and Generosity) but this time excluding those variables seen to be economic or government-related in nature (GDP Log and Perceptions of Corruption).

Support Vector Machine Model

Support Vector Machine (SVM) is another supervised learning algorithm that can be used to perform happiness prediction. SVM can also accomplish both regression and classification tasks, but for the purpose of this analysis, it will be used for classification so the prediction outcomes can be compared to other supervised learning efforts in this report. Similar to Naïve Bayes and Decision Tree, the objective is to see if Support Vector Machine can classify different levels of happiness (high, medium, low) based on values of the 6 supporting factors: GDP, healthy life expectancy, social support, generosity, freedom of life choices, and corruption perception.

A good first step to SVM analysis is to understand the linear separability of the data. Utilizing the top 2 most correlated features for happiness, GDP and healthy life expectancy, the classes for the training dataset (years 2013 to 2017) and test dataset (year 2018) are charted side by side. This visual analysis helps ensure that training and testing data are not too disparate.

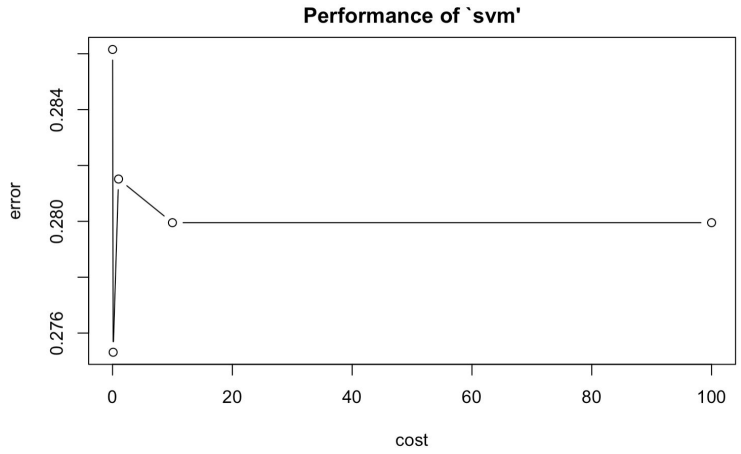


Given the multiple dimensionality presented by having 6 predictor features, 3 different SVM kernels are fitted with the training data to see which one would yield better results. SVM is also sensitive to the choice of parameters and small parameters changes may lead to very different classification results. Therefore, prior to fitting the model with different kernels, the optimization parameters are found by running the tune() function. Using a 10-fold cross validation, different parameter values are examined to see which set minimizes classification error of each kernel model. The parameters examined are cost of constraints violation (parameter C) using values .01, 0.1, 1, 10, and 100 and gamma using values 0.25, 0.5, 1, 2, and 3. Parameter C adjusts how hard or soft the classification margin should be (penalty for misclassified points), and the gamma parameter defines an overall scale factor for the distance between two points (with low values meaning “far” and high values meaning “close”), which in turn defines how a support vector shapes the decision boundary. However, a higher value of gamma will try to fit the training data too closely and cause an overfitting problem. In addition, for the polynomial kernel SVM, the number of degrees between 1 to 5 are also tested.

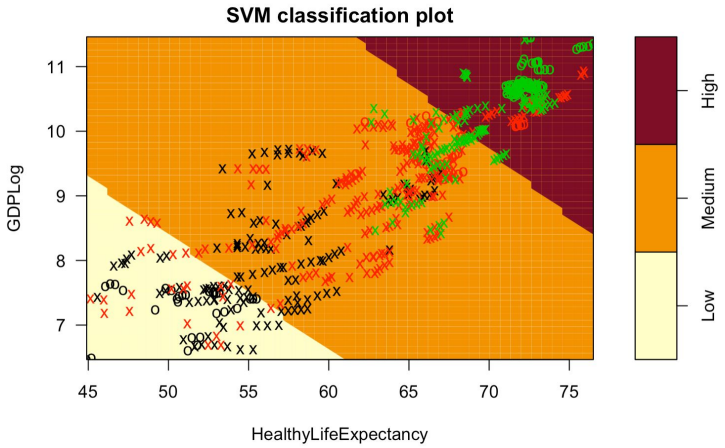
SVM with Linear Kernel

Here are configurations for the first SVM classification model fitted with the linear kernel. Using grid search, a set of best parameters is obtained and used in the model training.

- Kernel = linear
- Response = Happiness level (high, medium, low)
- Features = All 6 happiness supporting factors (GDP, healthy life expectancy, social support, generosity, freedom of choices, and corruption perception)
- Cost = 0.01 (from 0.01, 0.1, 1, 10, 100)
- Scale =True



A low C parameter yields the least error.

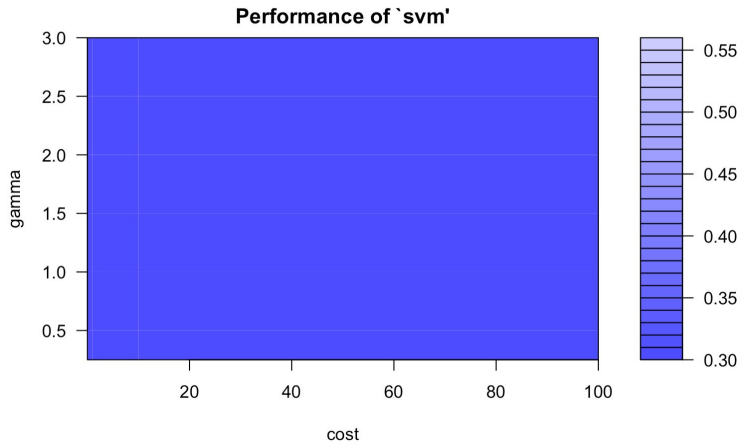


The linear SVM model plot of the training data indicates that it uses lines to separate the 3 classes of happiness. Since the correlation between GDP and healthy life expectancy is positive, the model makes 3 cuts to the data. Many data points in the training data are not easily separable by straight lines leading to a high likelihood of classification errors.

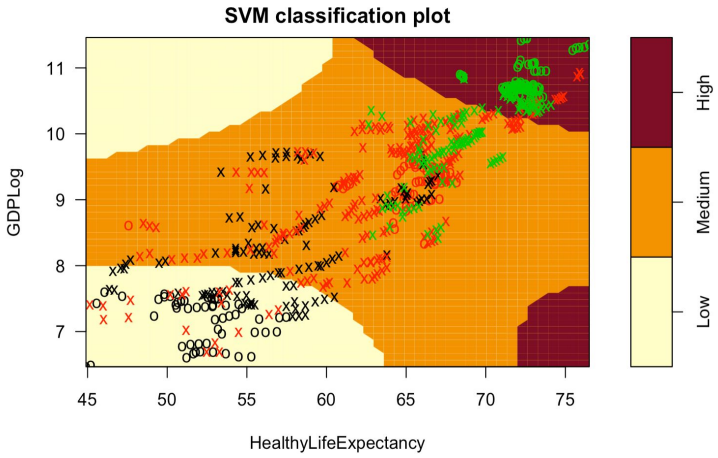
SVM with Polynomial Kernel

The second SVM model is fitted with the polynomial kernel. Using the tuning function, a set of best parameters is obtained and used in the model fitting. Below are the optimized parameters chosen for the model training.

- Kernel = polynomial
- Response = Happiness level (high, medium, low)
- Features = All 6 happiness supporting factors (GDP, healthy life expectancy, social support, generosity, freedom of choices, and corruption perception)
- Cost = 0.01 (from 0.01, 0.1, 1, 10, 100)
- Gamma = 1.5 (from 0.25, 0.5, 1, 1.5, 2, 2.5, 3)
- Degree = 3 (from 1 to 5)
- Scale = True



Grid search recommends a low C parameter with gamma of 1.5.

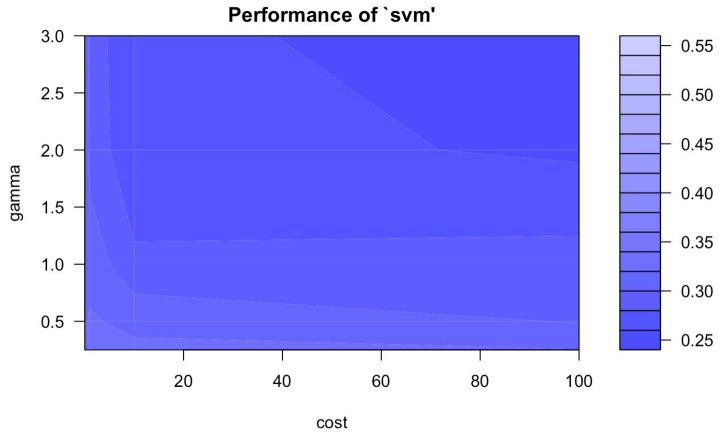


The polynomial kernel generates new features by applying polynomial combination of all existing features and thus allows for learning of non-linear models. Using these polynomial features, this SVM model generates non-linear decision boundaries. When compared to the linear kernel output, this model is able to separate out more of the medium happiness observations from the low and high data points. But in doing so, more observations for low or high happiness would also likely to be misclassified as medium.

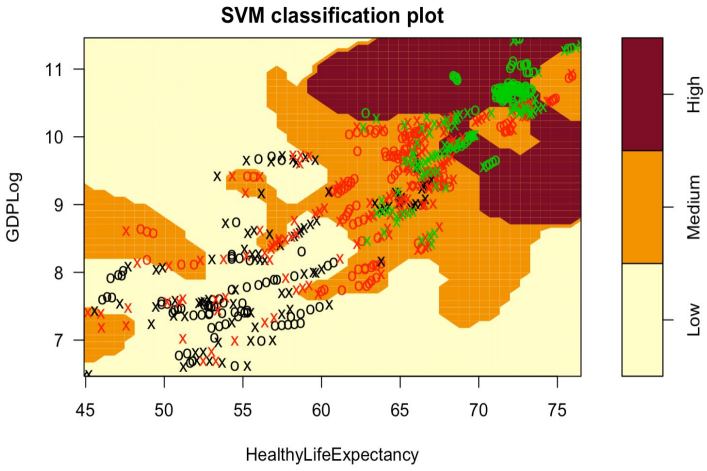
SVM with Radial Basis Kernel

The last SVM analysis is implemented using the radial basis kernel. Again, C and gamma hyper-parameters are obtained using grid search. Below are the optimized model parameters used for model training.

- Kernel = radial basis
- Response = Happiness level (high, medium, low)
- Features = Top 2 highest happiness supporting factors (GDP and healthy life expectancy)
- Cost = 100 (from 0.01, 0.1, 1, 10, 100)
- Gamma = 3 (from 0.25, 0.5, 1, 2, 3)
- Scale = True



High C and gamma parameters would yield the least error.



Radial kernel SVM is a more robust approach to handle happiness data points that are not linearly separable. The attempt of this iteration is to improve prediction outcomes. First, the model is simplified by involving only 2 input features (GDP and healthy life expectancy). Second, the radial kernel trick if fitted with higher values of C and gamma. A high gamma value is able to create boundaries around data points of different classes of close proximity, where the “curve” of the decision boundaries is visible in the SVM plot. A larger C parameter tries for better accuracy and makes the classifier avoid misclassified data points by penalizing for misclassified data.

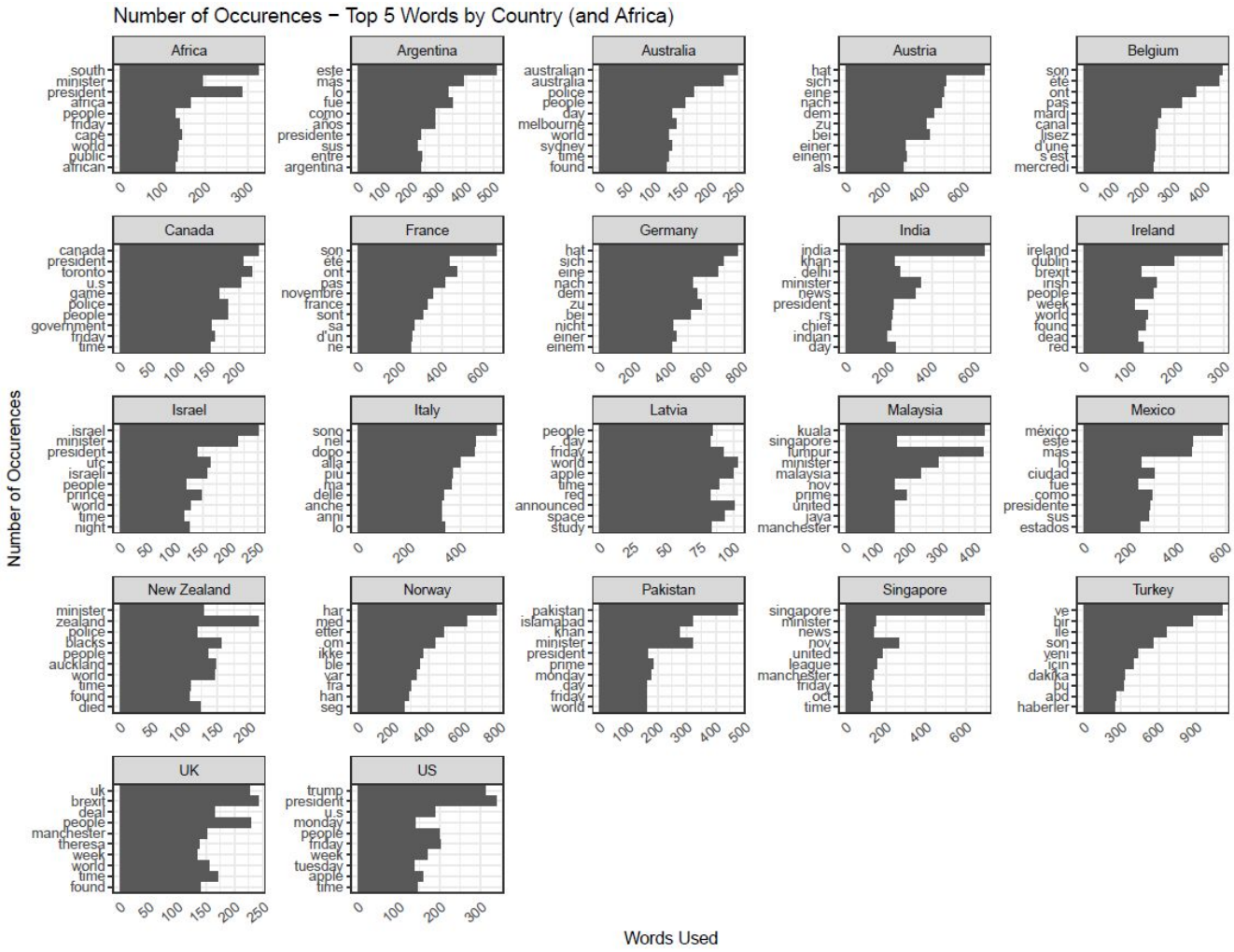
Text Mining Model

As outlined in the World Happiness Report, countries with higher levels of governmental corruption, poor living conditions, and those experiencing turmoil such as conflict or war, will logically tend to have lower scores among the various elements the report measures. To explore this trend, a dataset of one month's worth of news headlines from countries across the world is analyzed.

This dataset was created using Google News aggregation; it contains approximately 93,000 individual headlines, gathered from 23 different countries and one world region, Africa. The headlines themselves vary in language; unfortunately, Russian and Chinese headlines are indecipherable in the dataset, as these headlines are stored in unreadable characters, most likely due to the nature of Chinese script and the Cyrllic alphabet requiring specialized font packages and/or unknown text translators. Thus, to preserve the integrity of the dataset, all Chinese and Russian headlines are removed from the headlines. This leaves 21 countries and one region all having headlines for analysis. French headlines are left intact, as the authors were able to translate sufficiently with their language skills.

Firstly, to prepare all of the headlines for analysis, the entire headline dataset is tokenized in R using the TM package. After tokenization, all of the tokenized words are analyzed to reveal the most commonly used words among all of the 21 countries. Unsurprisingly, the most common words across all languages appear to be stopwords, including "the," "and," and so forth. To facilitate in-depth analysis, these stopwords are removed, across all languages.

After removing the stopwords, a visualization of each country and region's top five most commonly used words is created using ggplot. This chart, which follows below, reveals interesting commonalities and disparities among the countries and region:

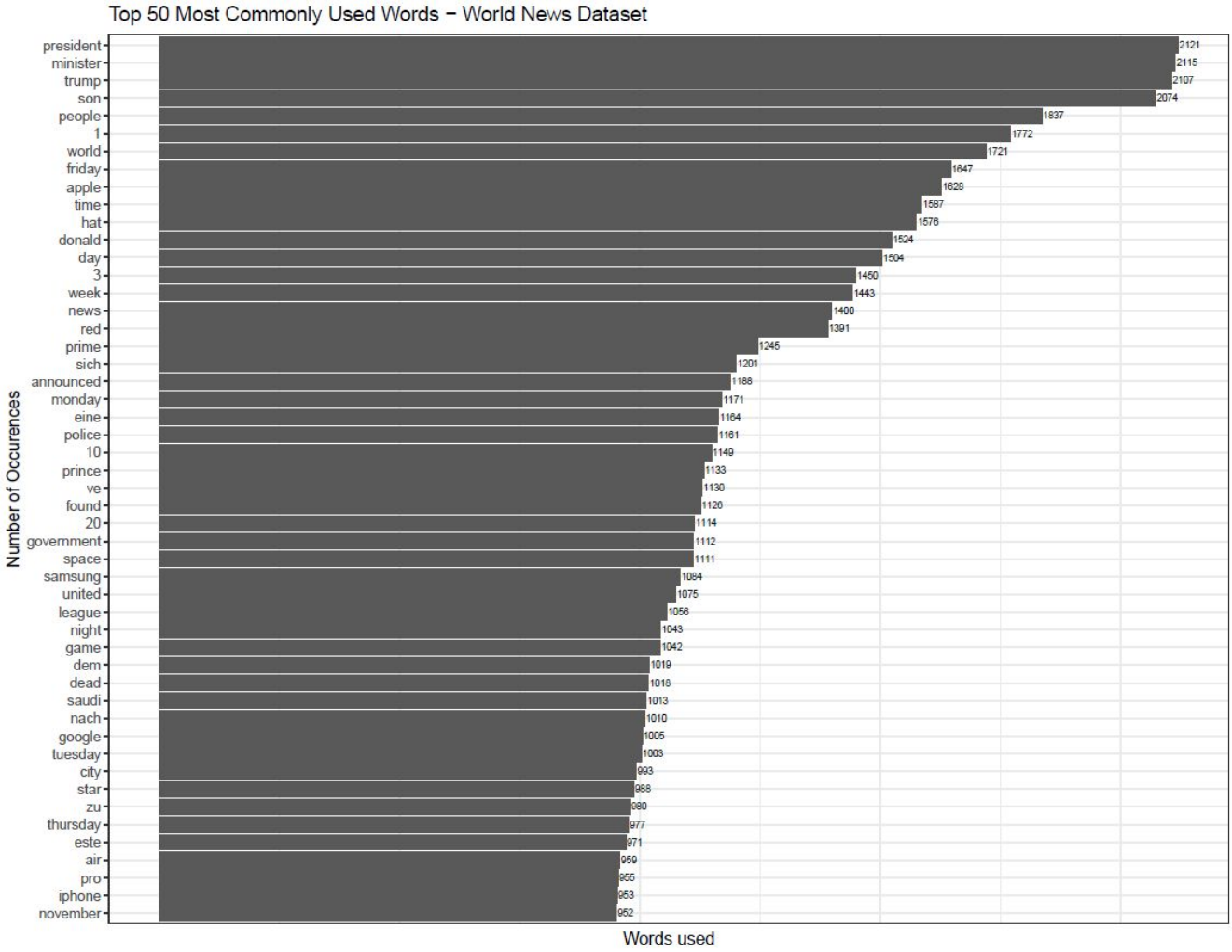


Above: A visualization showing the top five most commonly used words across all headlines for Africa and all countries in the dataset, broken down by country.

Many of the countries' top five words contain the country's name itself. Self-referential word use reveals that one can surmise each country is concerned primarily with news reflecting its own affairs and within its borders. There are instances though, for example, that of Canada, where important neighboring countries appear among these words -- Canada's top five words contain "U.S.", referring to the United States; this appears to emphasize that relationship that the USA and Canada is quite important to each respective country. Future analysis should examine these patterns in greater detail, ideally adding more countries and incorporating Association Rules Mining and other techniques to delve further into hidden news patterns.

After this exploratory analysis of the words in the dataset, further analysis with visualization in ggplot highlights more patterns among words and word use among the headlines.

An initial visualization of the top 50 most frequently used words across all headlines reveals interestingly that “2018” and “2” were the most commonly used; after removing these, since they provided little context, the visualization is recreated, and follows below:



Above: a bar graph displaying the top 50 most frequently used words in all countries’ and region’s headlines.

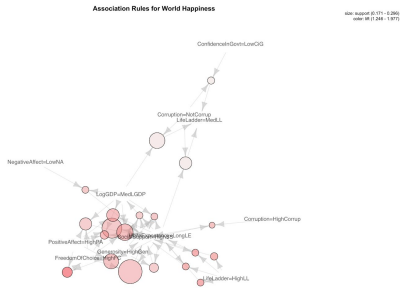
This chart demonstrates the extent to which the United State’s headlines and politics dominate headlines not only in the USA but across many world regions. With “president” and “trump” being among the top five words, and “Donald” being among the top 25, there is little doubt that the USA’s political events are covered heavily across the world. It is also interesting that days of the week occur so commonly in the headlines; this shows that among all of the countries and region of Africa, people are very concerned with when events occur, and when they will occur. Two large corporations, Google and Apple, also occur quite often in these headlines, along with “iphone”. This seems to be an indication of how pervasive technology is in all of these countries and region of Africa.

We then turn to sentiment analysis, the results of which are discussed below in Results.

RESULTS

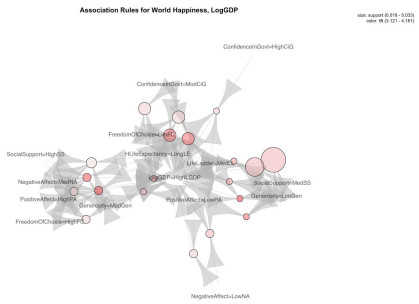
Association Rule Mining

Running the association rules resulted in several key insights:



The combination of a long life-expectancy with a strong sense of freedom of choice or strong positive affect had a strong association with a high sense of social support. Countries with a strong sense of freedom of choice additionally had a strong sense of generosity and positive affect. Notably, if a country had a moderate life ladder estimate, there is a strong association with low confidence in government, though not necessarily an indication that there is corruption. High log of GDP per capita is absent from the strongest association rules.

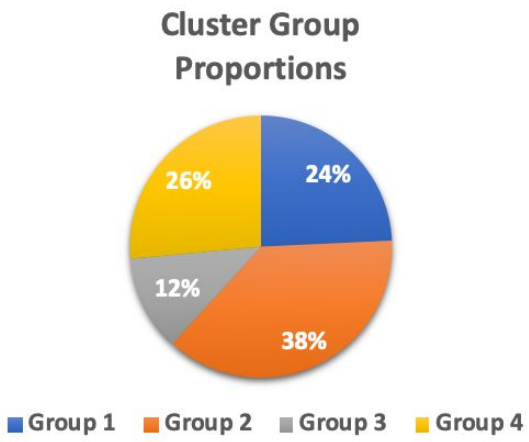
When creating rules that account for a high log of GDP per capita, several trends are notable:



Countries with a high log of GDP per capita often demonstrated moderate levels of social support and limited sense of generosity. Additionally, the combination of a mid-grade average of a country's estimation of their place on the life ladder, a long healthy life expectancy, limited freedom of choice, low positive affect, and moderate confidence in government build a narrative of countries that have wealth and live long lives, but are only moderately happy or satisfied. Rule 9 in the figure below builds a more positive narrative when considering the addition of a strong sense of social support and high freedom of choice.

When modeling for high life ladder values on the right-hand side, the strongest rules illustrated the highest potential for HighLL when social support, healthy life expectancy, and lots of freedom or choice were all present in a country. The rules that included GDP at all had medium levels of GDP at best. This suggests that even if countries are not the wealthiest, so long as they have a moderate amount of wealth and strong social systems, it was possible to achieve a high level of satisfaction.

K-means Clustering



K-means cluster analysis identifies 4 segments within the 2018 data, with a fairly even split.

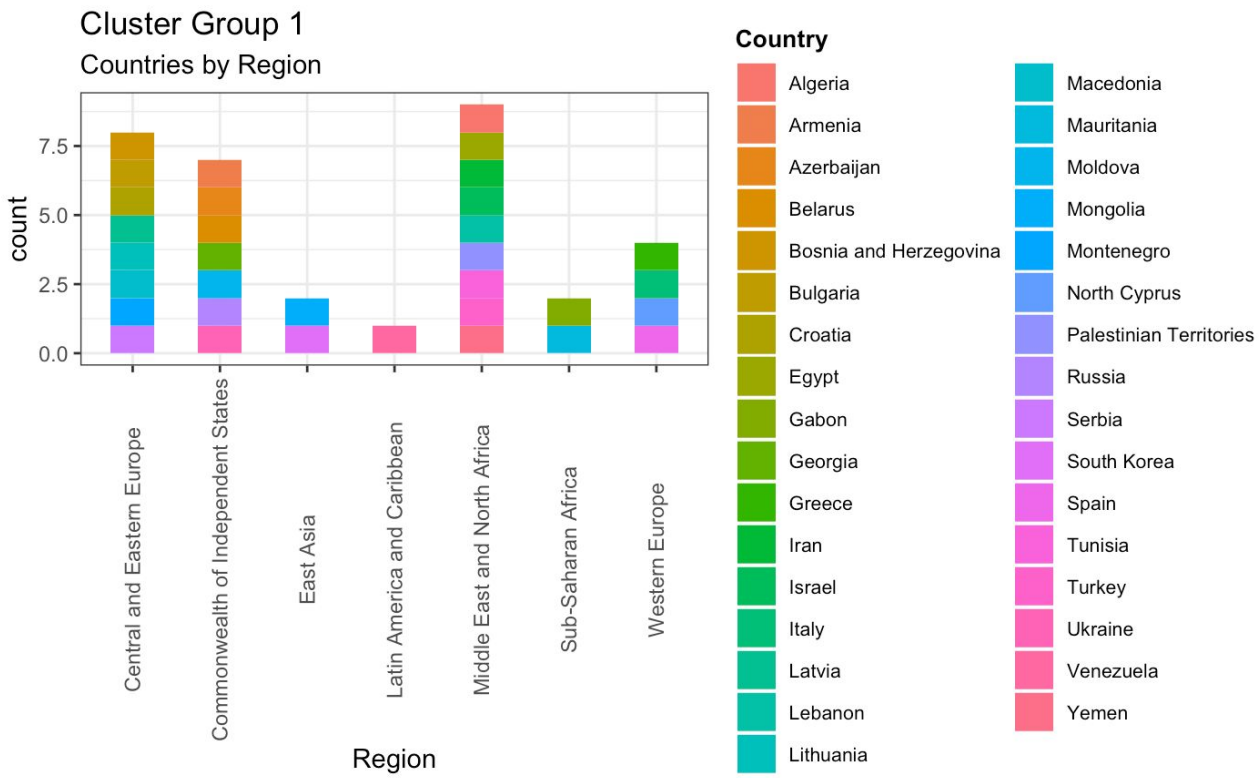
Figure 14: K-means cluster assignments

1	2	3	4
33	51	16	36

K-means cluster segment profiles

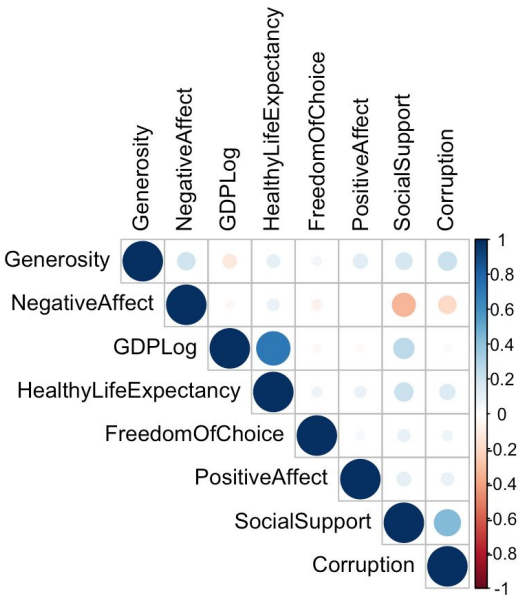
In order to highlight similarities - and differences - between each cluster assignment as identified using K-means cluster analysis, a correlation matrix was run for each segment to highlight any specific correlation patterns and see how these compare and contrast between each group.

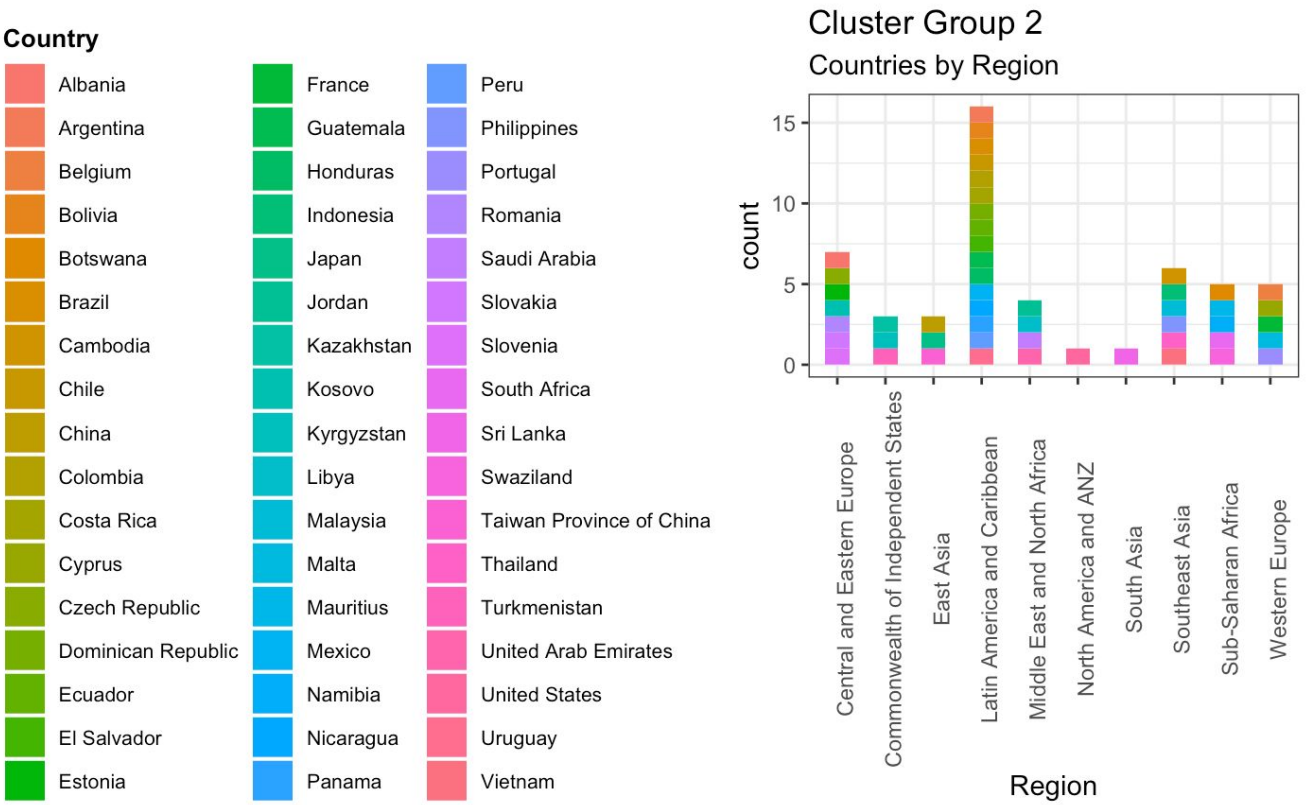
The following matrices for each group highlight distinct differences between the correlation of specific indicator variables within those regions. For example, when looking at cluster segment 3 the most highly negatively correlated values appear to be Social Support and Corruption. According to the countries by Region chart, this segment features Western European countries most heavily. This finding therefore indicates that within these regions, a higher prevalence of social support systems results in lower perceptions of corruption within that nation's government. In addition, when looking at the country/region charts for each cluster, it is interesting to see that on the basis of the predictor variables that are used, groups appear to be clustered generally along regional lines with each segment featuring one region most heavily. This provides an ability to profile each cluster segment according to the region(s) it features, and the relationship between variables as highlighted by the correlation matrix.



Key observation: Cluster 1 features countries in the **Middle East and North Africa** most heavily.

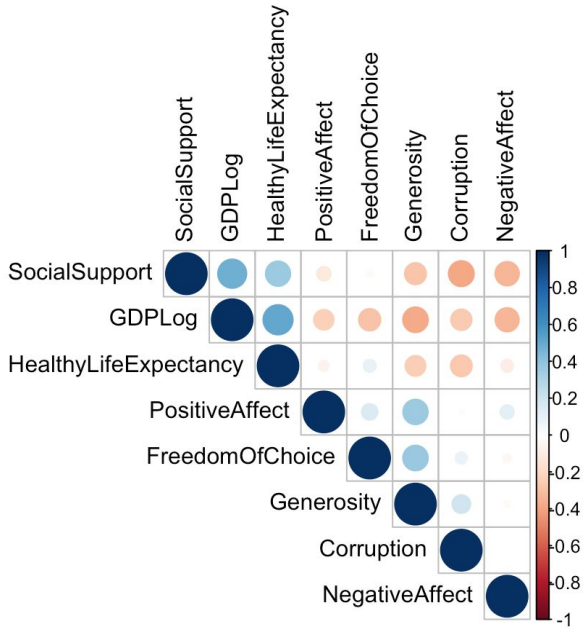
Generalizations: in these Middle Eastern and North African countries, there is a strong negative relationship between social support systems and negative affect. The prevalence of social systems (or lack thereof) clearly affects citizens in a negative way.

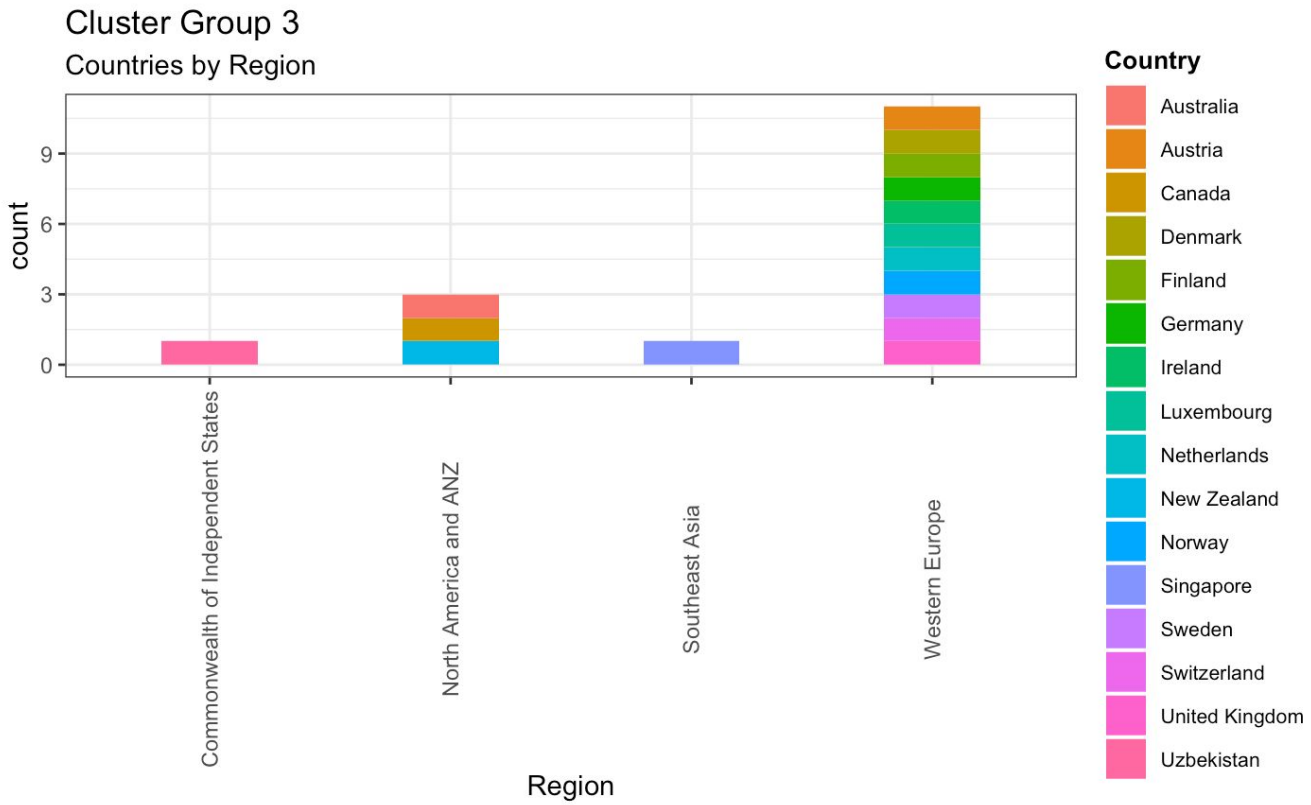




Key observation: Cluster 2 features countries in **Latin America and Caribbean** most heavily.

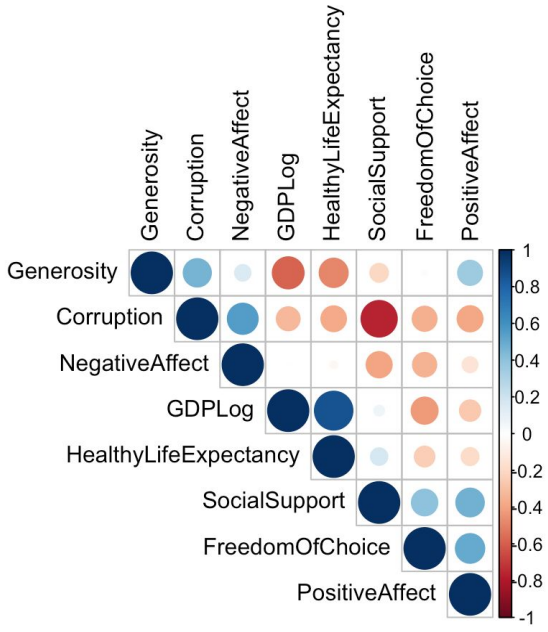
Generalizations: in these Latin American and Caribbean countries, a higher prevalence of social support systems results in lower perceptions of corruption within each nation's government.



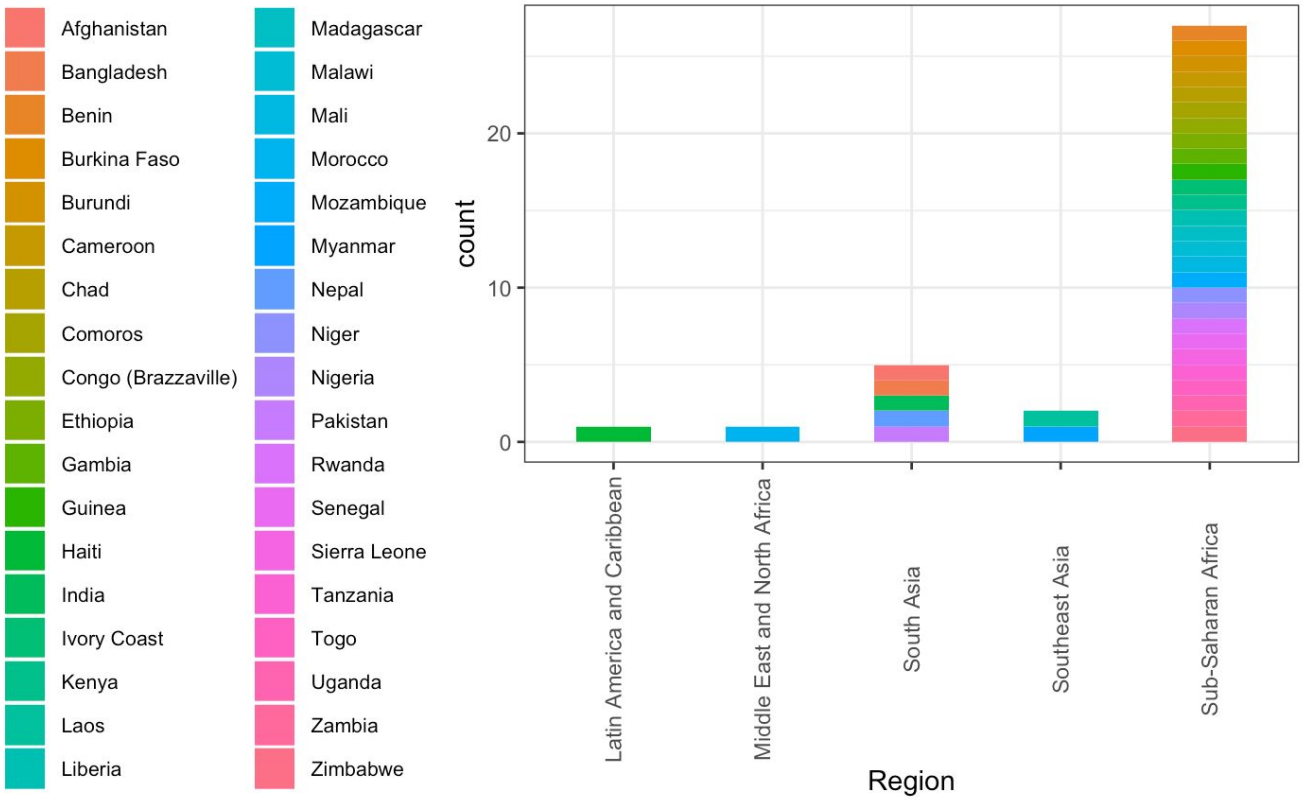


Key observation: This group features countries in **Western Europe** most heavily.

Generalizations: in these Western European countries, a higher prevalence of social support systems results in lower perceptions of corruption within each nation's government.

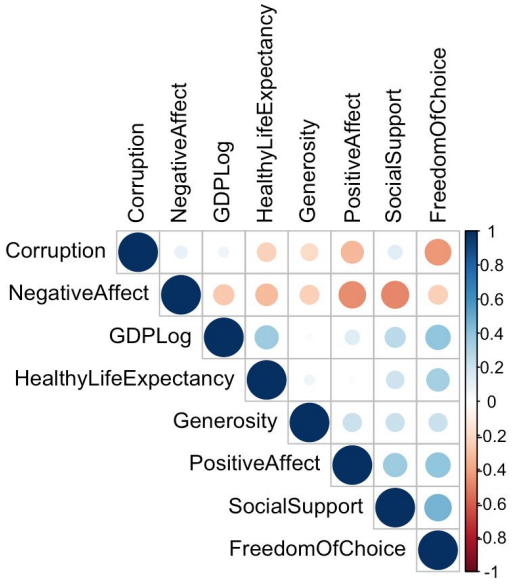


Cluster Group 4 Countries by Region



Key observation: This group features countries in **Sub-Saharan Africa** most heavily.

Generalizations: *in Sub-Saharan African countries a high negative correlation exists between Freedom of Choice and Perception of Corruption; also Social Support and Negative Affect. There is however a strong **positive** link between Freedom of Choice and Social Support in these countries.*



Naive Bayes

The initial run using 2 prediction scores ("unhappy" and "happy") results in a fairly high accuracy score. Taking the largest number of previous years into account (2013-2017) the confusion matrix shows as follows:

Naive Bayes validation set (2013-2017) confusion matrix: 2 categories

	ACTUAL	
PREDICTION	**UNHAPPY**	**HAPPY**
UNHAPPY	**6**	7
HAPPY	1	**51**

Accuracy: **88%**

Results for 2018 only were slightly less precise but still show a good degree of accuracy:

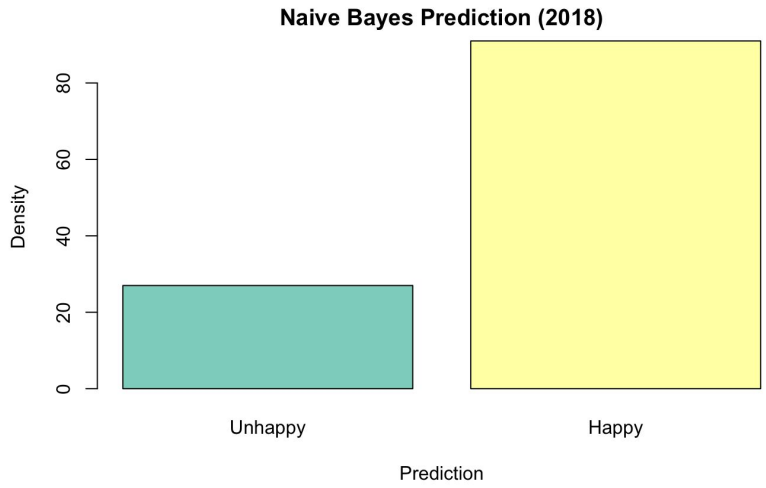
Naive Bayes 2018 confusion matrix

	ACTUAL	
PREDICTION	**UNHAPPY**	**HAPPY**
UNHAPPY	**7**	20
HAPPY	3	**88**

Accuracy: **81%**

When using this method to predict between "unhappy" and "happy", there remains a reasonably large discrepancy between the results of the two categories. It is at least encouraging to see that the happy category is a much denser group!

Density spread of Naive Bayes 2018 prediction (2 categories)



Interestingly, results obtained when select variables were used rather than all available predictors - regardless of combination - showed a less accurate prediction model when applied to the validation set (83%) but a comparatively more accurate one (86%) when applied to the holdout 2018 data. This appeared to hold true regardless of variable combination.

In predicting three categories of class label (“Low”, “Medium” and “High) the Naive Bayes algorithm proved less successful overall in terms of accuracy in prediction. With a larger number of possibilities to predict, it appears there is a larger margin of error when comparing the original happiness scores against the discretized levels used for prediction by the model.

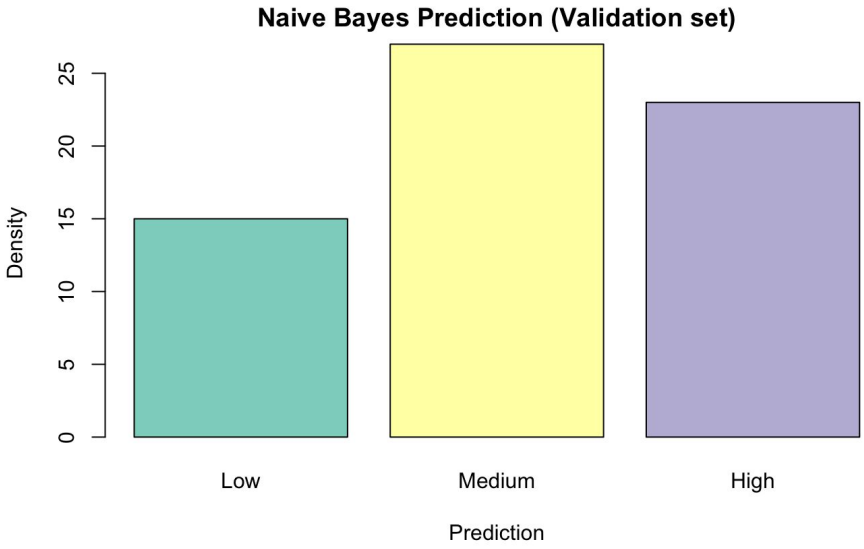
Naive Bayes validation set (2013-2017) confusion matrix: 3 categories

	ACTUAL		
PREDICTION	**LOW**	**MED**	**HIGH**
LOW	**10**	**5**	**0**
MEDIUM	**4**	**20**	**3**
HIGH	**0**	**3**	**20**
	Accuracy:	***77%***	

Naive Bayes 2018 prediction confusion matrix: 3 categories

	ACTUAL		
PREDICTION	**LOW**	**MED**	**HIGH**
LOW	**19**	**12**	**0**
MEDIUM	**4**	**38**	**7**
HIGH	**1**	**6**	**31**
	Accuracy:	***75%***	

Density spread of Naive Bayes 2018 prediction (3 categories)

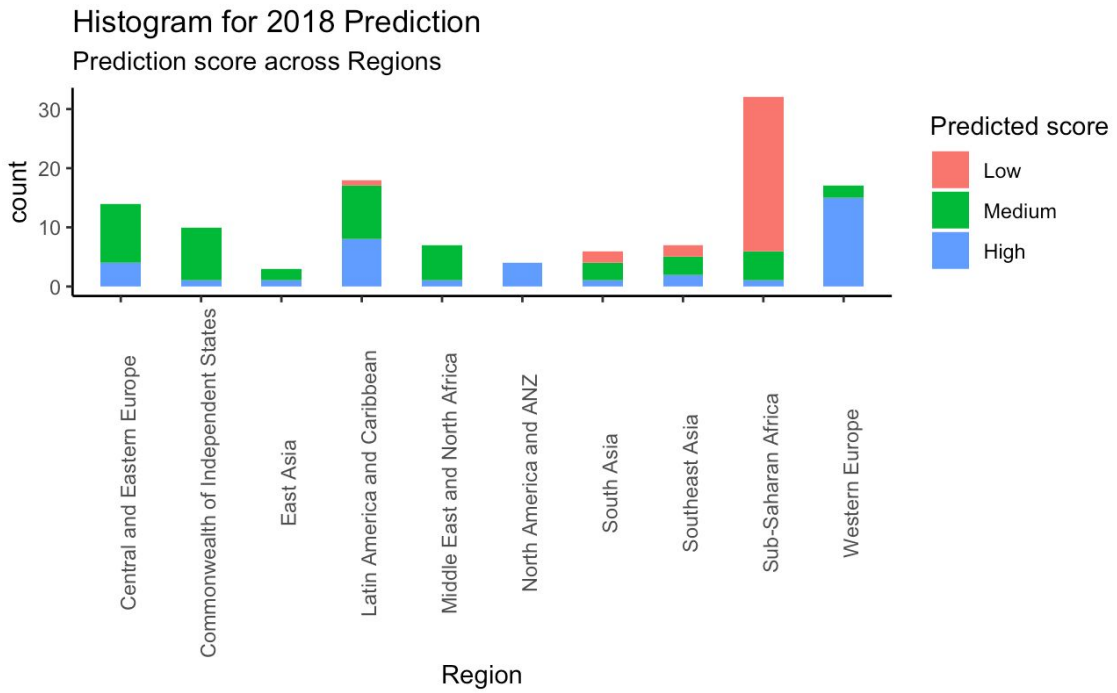


When predicting three categories, there is a much more even split between those countries predicted to be “less happy” or “most happy”, with the majority being clumped into the medium category. This is likely to be more representative of real-life!

Detailed results by country and region using Naive Bayes prediction method

Delving more deeply into the results obtained by country, the prediction scores can be identified by region in order to provide an overview:

Naive Bayes 2018 Prediction Score by Region



To differentiate the happiest countries from the least happiest, both the 'Top 10' and 'Bottom 10' countries can be extrapolated in order to see if it is possible to identify any differences.

Top 10 Happiest Countries according to Naive Bayes prediction

By highlighting the largest value for each predictor variable in the top 10, it is clear the top 5 happiest countries also contain the highest value scores for social support, healthy life expectancy, and freedom of choice - but **NOT GDP**.

Top 10	GDP	Social Support	Healthy Life Expectancy	Freedom of choice
Finland	**10.64**	**0.96**	**71.90**	**0.94**
Denmark	**10.76**	**0.96**	**72.40**	**0.94**
Switzerland	**10.98**	**0.93**	**74.10**	**0.93**

Netherlands	**10.81**	**0.94**	**72.30**	**0.92**
Norway	**11.09**	**0.97**	**73.20**	**0.96**
Austria	**10.74**	**0.91**	**73.00**	**0.90**
Sweden	**10.77**	**0.93**	**72.60**	**0.94**
New Zealand	**10.50**	**0.95**	**73.20**	**0.95**
Luxembourg	**11.45**	**0.90**	**72.60**	**0.88**
United Kingdom	**10.60**	**0.93**	**72.30**	**0.84**

Bottom 10 Least Happiest Countries according to Naive Bayes prediction

Similarly, when looking at the 5 least happiest countries, this category (specifically **Afghanistan** which is officially the least happiest country of all) also has the lowest value score for healthy life expectancy and freedom of choice, but **not lowest GDP** score.

Bottom 10	GDP	Social Support	HLE	Freedom of choice
Zambia	**8.22**	**0.72**	**55.30**	**0.79**
Togo	**7.29**	**0.60**	**54.70**	**0.61**
Comoros	**7.26**	**0.62**	**57.20**	**0.56**
Burundi	**6.54**	**0.48**	**53.40**	**0.65**
Zimbabwe	**7.55**	**0.78**	**55.60**	**0.76**
Haiti	**7.42**	**0.54**	**55.70**	**0.59**
Rwanda	**7.57**	**0.62**	**61.10**	**0.92**
Tanzania	**7.93**	**0.68**	**57.50**	**0.81**
Malawi	**7.01**	**0.53**	**57.60**	**0.80**
Afghanistan	**7.49**	**0.51**	**52.60**	**0.37**

These results appear to indicate that high or low GDP is not necessarily an indication of how happy a country will actually be. In contrast, a longer healthy life expectancy and the freedom to make one's own decisions in life appear to be a much higher indication of happiness.

Prediction Results and social factors

To further explore feature differences between the least, medium, and happiest countries (and their associated regions), it is interesting to compare the values and spread of the more socially-oriented variables.

Average Healthy Life Expectancy per Predicted Happiness Category

	Average
LEAST HAPPY	55.28
MODERATELY HAPPY	65.56
HAPPIEST	70.88

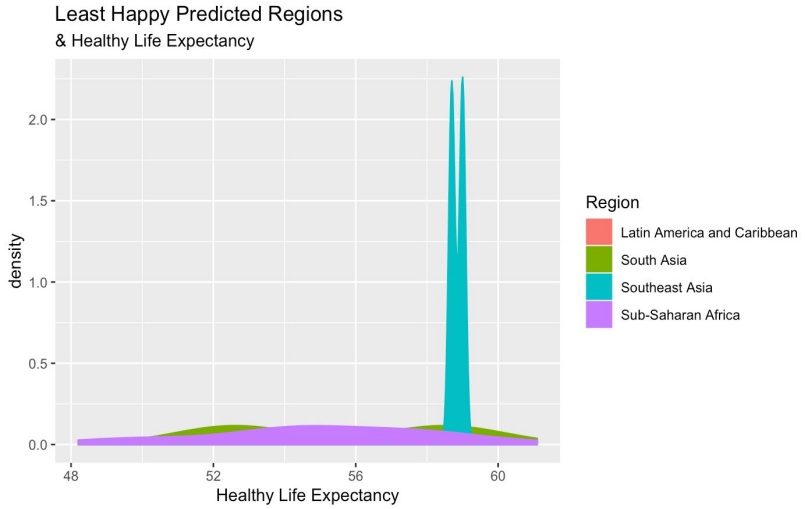
As may be expected, countries that are designated as the least happy have a relatively low average life expectancy at birth, whereas countries that are happiest enjoy a much greater degree of longevity. Levels of social support also show a relatively high differentiation between each group, as evidenced by the average values in each:

Social Support levels per Predicted Happiness Category

	Average
LEAST HAPPY	65
MODERATELY HAPPY	82
HAPPIEST	91

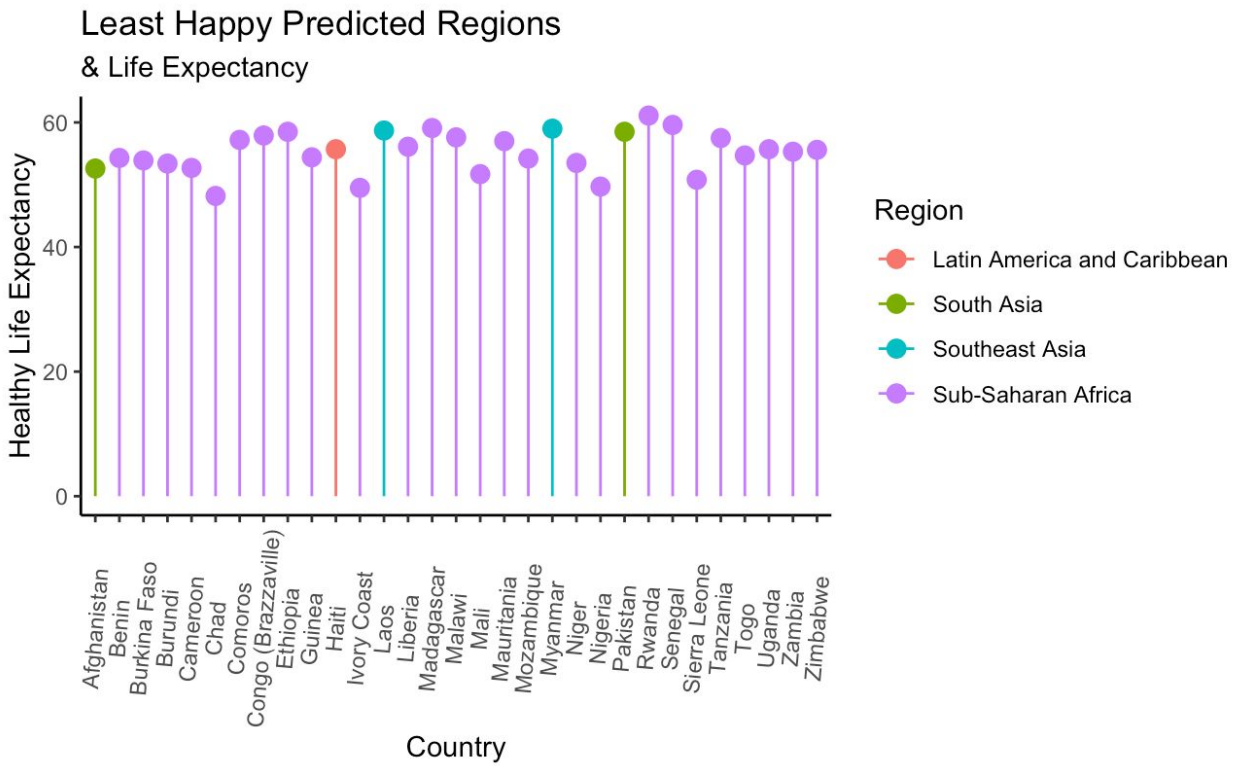
Individual variables may now be looked at in greater detail with regards to each prediction group, in order to see if any key takeaways can be identified.

Least Happy: Healthy Life Expectancy by Region



Countries within Sub-Saharan Africa show a very wide spread of life expectancy values within the countries that make up this region. As Laos and Myanmar both have a relatively high HLE (58.7 and 59 respectively) then S.E. Asia shows a density spike.

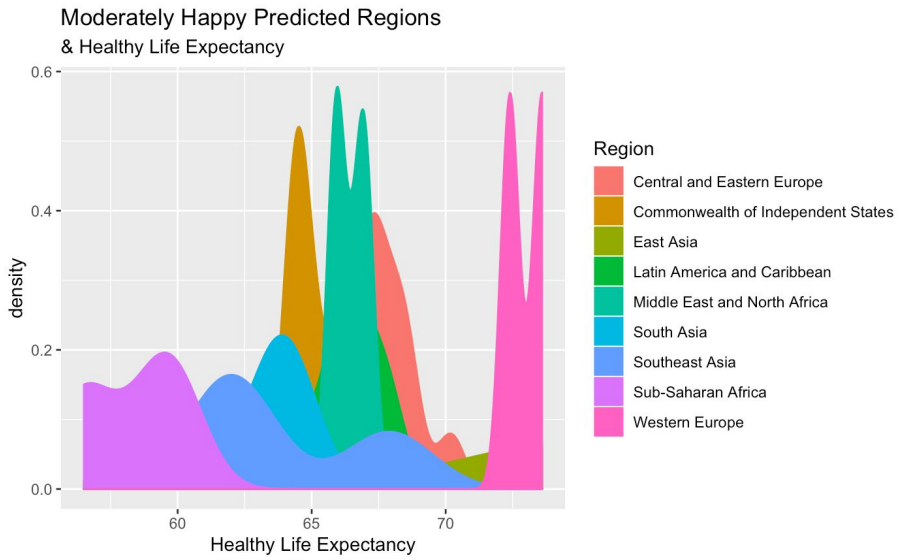
Least Happy: Healthy Life Expectancy by country



Here we see that, as shown in the table of averages, the countries predicted to be the least happy generally have a healthy life expectancy between 50 - 60 years of age. Only Rwanda appears to have a life expectancy over 60 years (61).

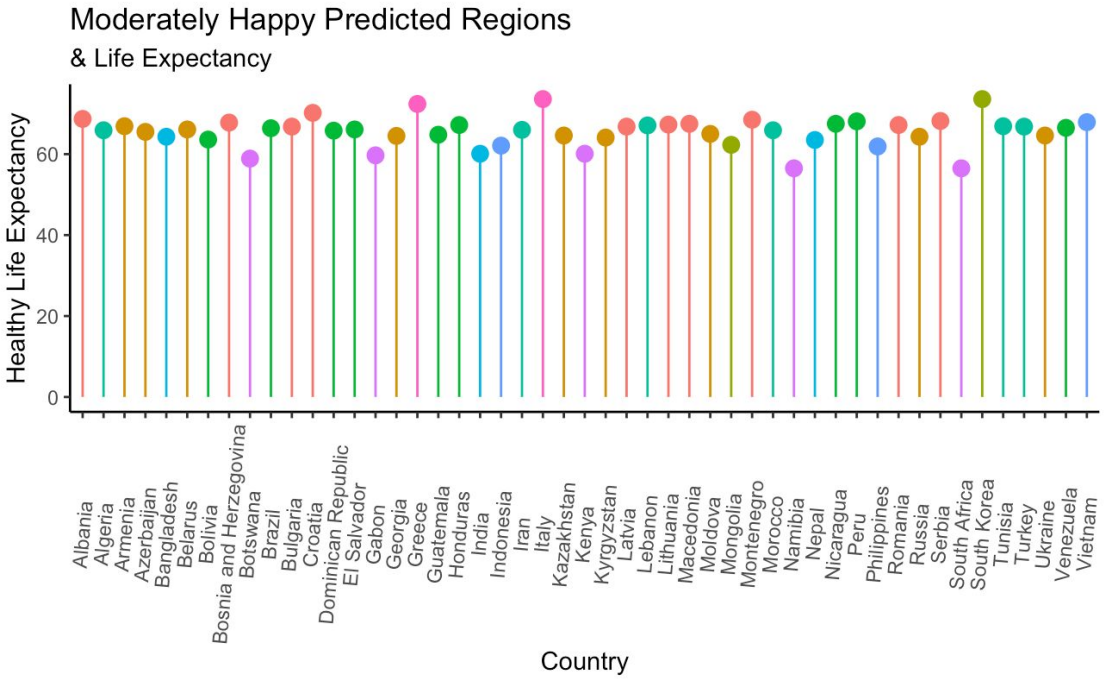
In contrast, those countries appearing in the ‘moderately’ happy category have a healthy life expectancy of 60 years and above. The majority of countries in this group do not exceed 70 years, with the exception of Italy, South Korea and Greece, which have life expectancies of 72.4 (Italy) and 73.6 (S. Korea and Greece) respectively.

Healthy Life Expectancy by Region

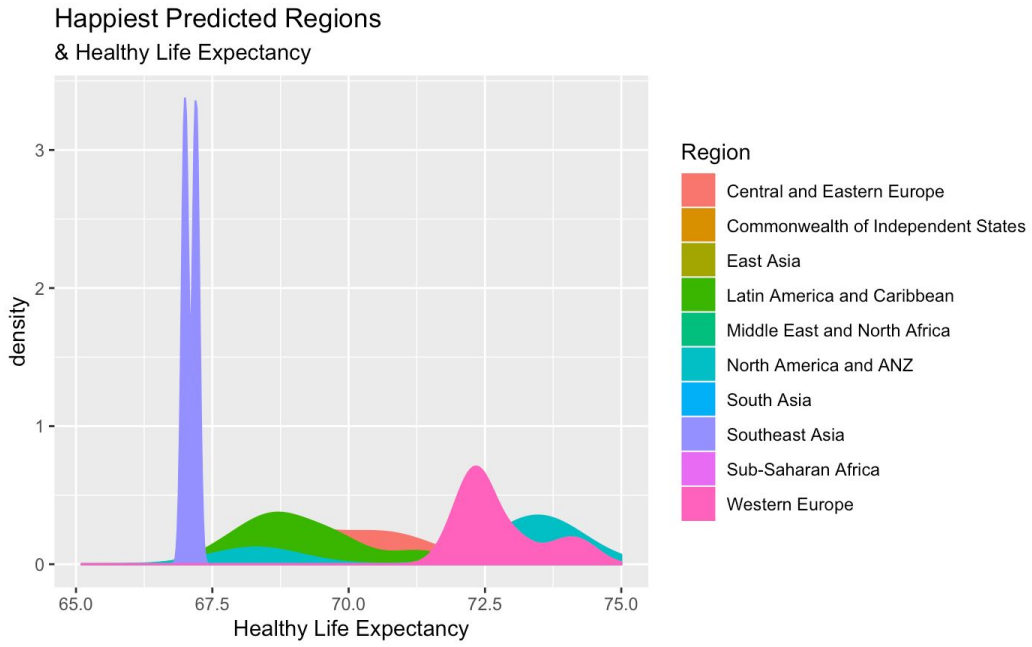


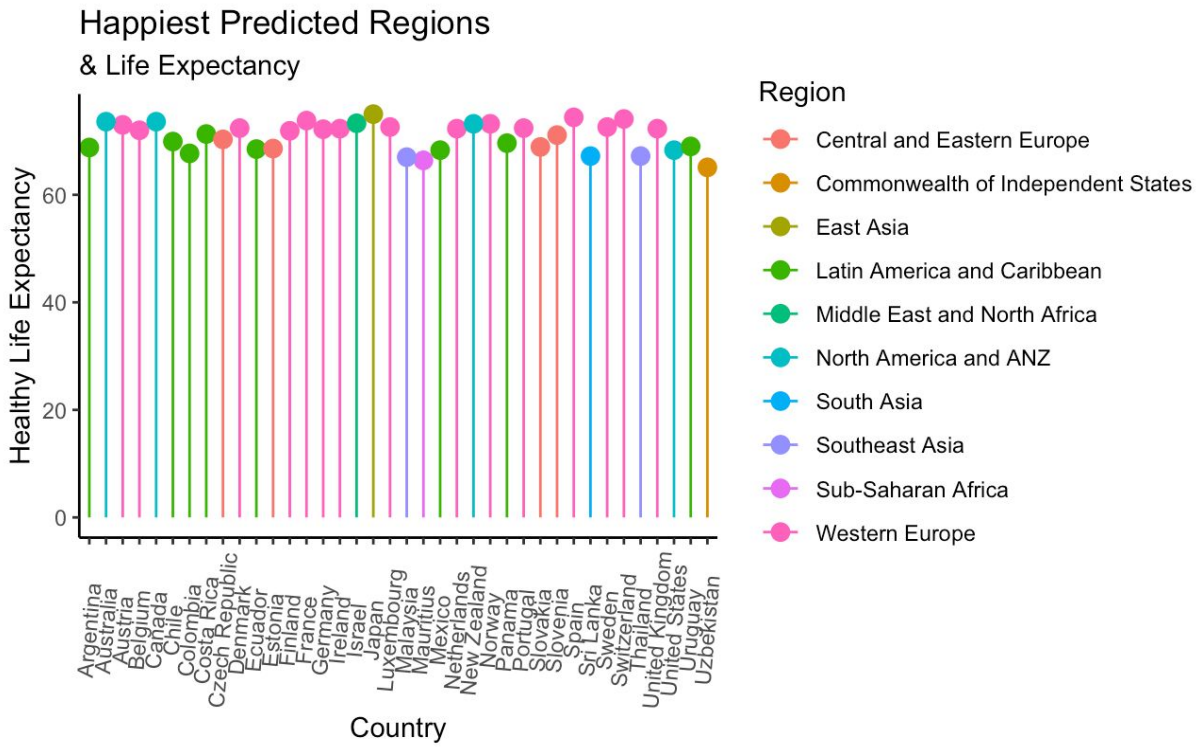
Moderately happy regions have the most variation regarding life expectancy within countries. Western Europe features the highest density of countries with the highest life expectancy, with **Greece** and **Italy** having the highest life expectancy within those western European countries predicted to be of moderate happiness.

Healthy Life Expectancy by country



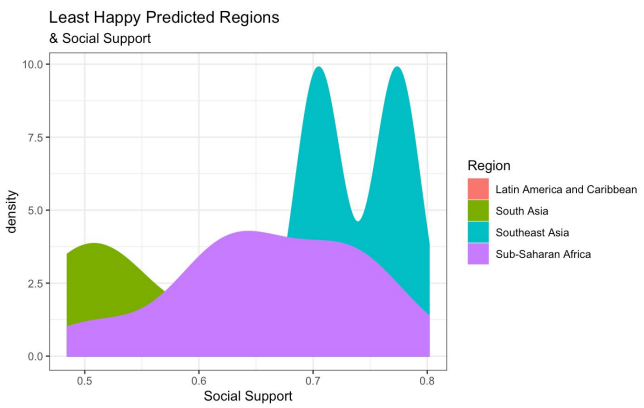
Healthy Life Expectancy by Region

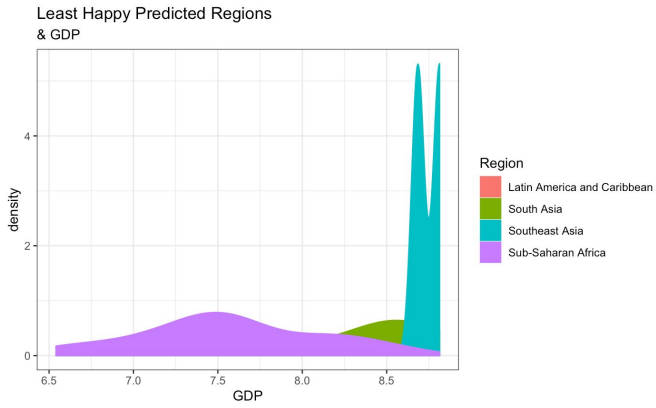




Within the happiest countries group, it is reassuring to see that Denmark features within our prediction! Referring to the prior observations regarding the U.S. and the U.K. it is interesting to note here that the U.K. has a slightly higher Healthy Life Expectancy. This obviously plays a defining factor in the "Top 10" happiest countries designation, as the GDP value of the U.K. is definitely far less than the United States.

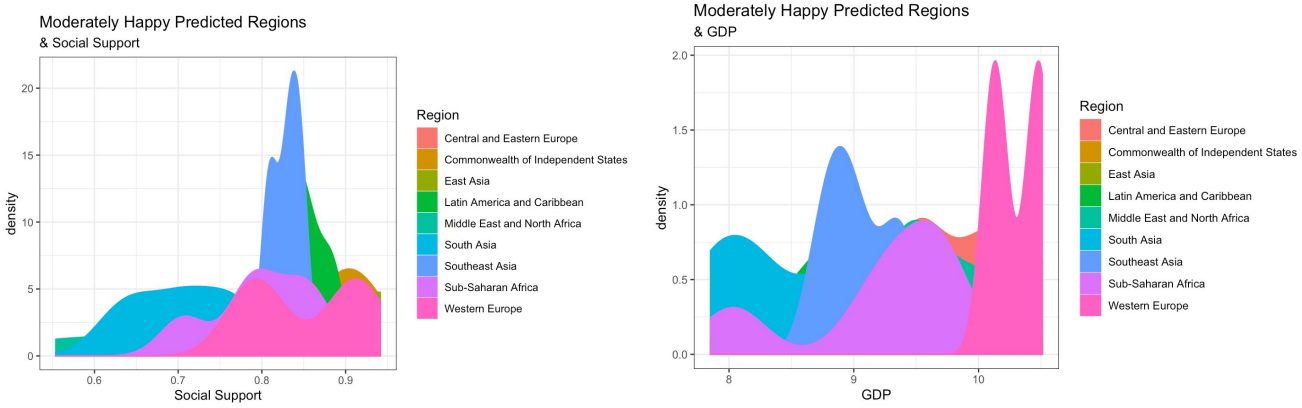
GDP and Social Support comparison



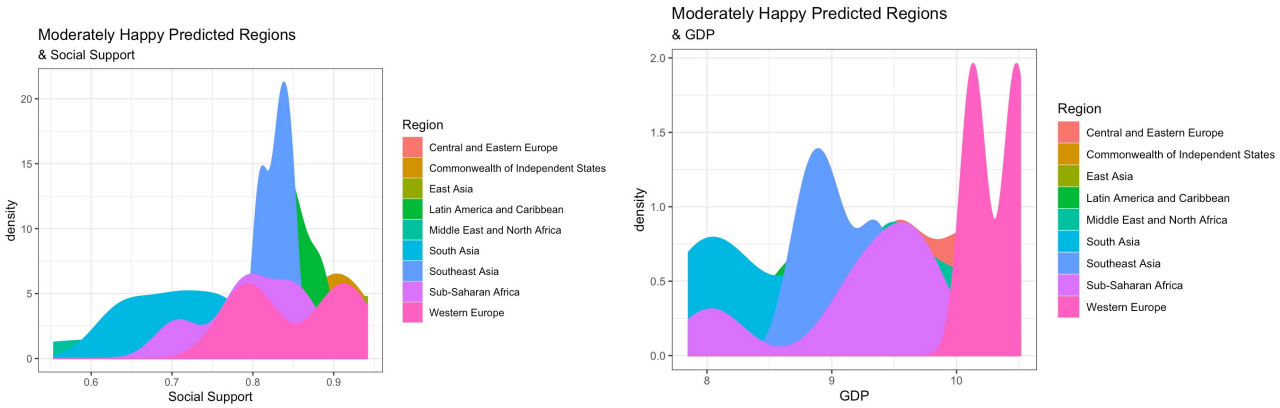


In comparing economic factors (GDP) with social factors (Social Support), what is revealing with respect to the “least happy” group is that the South Asia region has among the highest levels of GDP and yet values for social support are among the lowest. This would indicate that low levels of social support, rather than a higher GDP within these countries is a better indicator of how citizens perceive their overall happiness levels.

GDP and Social Support: Moderately happy regions



Again higher levels of social support are apparent in the Southeast Asia region compared to levels of GDP, which is on the lower-to-medium end in comparison to others in this same category.



GDP and Social Support: Happiest regions

Countries in Central and Eastern Europe here can be seen to have high levels of social support on par with countries in Western Europe and North America / Australasia, yet their GDP is lower. High social support factors however place them on an even keel with these countries in terms of overall happiness.

Decision Tree

After fitting the decision tree model with the training data, it is used to generate predictions for the test dataset, which contains predictor values for each country in the year 2018. Using a confusion matrix, the original 2018 happiness labels are compared with the model predictions, and the model reports an accuracy score around 72%.

The prediction output points to a major misclassification area where the decision tree predicts 14 countries with medium happiness level as having a low level of happiness. Upon further analysis of the tree's decision nodes, an issue is observed where the tree splits are highly contingent on healthy life expectancy and GDP. Given that this decision tree is more pruned and has less nodes, the model is unable to account for variation of happiness level in countries that have low healthy life expectancy and low GDP based on other 4 supporting factors. In many cases, countries exhibiting low healthy life expectancy and low GDP actually have a moderate level of happiness. However, the decision model generalized them as being the least happy.

Below is the confusion matrix when comparing the decision tree predictions with the original labels in the testing data and classification errors. Further tuning opportunities exist to create a less pruned decision tree that is capable of delivering more accurate results.

```
Confusion Matrix and Statistics

          Reference
Prediction Low Medium High
    Low     20     14    0
    Medium   4     37   10
    High     0      5   28

Overall Statistics

               Accuracy : 0.7203
                 95% CI : (0.6302, 0.799)
    No Information Rate : 0.4746
    P-Value [Acc > NIR] : 5.31e-08

                  Kappa : 0.5672

 Mcnemar's Test P-Value : NA
```

Model uses only 2 decision points when determining if a country happiness level is low.

Support Vector Machine

To determine accuracy of the 3 SVM models with different kernels on unseen data, each one is used to generate predictions for the test dataset, which contains 2018 happiness scores and their associating predictor values that are not part of the training dataset. After generating the prediction for each country, a confusion matrix is produced to compare the original happiness labels with the predicted ones. The confusion matrix calculates the number of accurate predictions as a percentage of the total number of observations and enables a quick side by side comparison of the SVM models.

Below are accuracy outputs for the linear SVM model- 73% (left), the polynomial kernel SVM model - 71% (center), and the radial basis SVM model 80% (right).

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.01

Number of Support Vectors:  513

Confusion Matrix and Statistics

          Reference
Prediction Low Medium High
    Low     18     10    0
    Medium   6     45   14
    High     0      1   24

Overall Statistics

               Accuracy : 0.7373
                 95% CI : (0.6483, 0.814)
    No Information Rate : 0.4746
    P-Value [Acc > NIR] : 5.872e-09

                  Kappa : 0.5777
```

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  0.01
     degree:  3
     coef.0:  0

Number of Support Vectors:  458

Confusion Matrix and Statistics

          Reference
Prediction Low Medium High
    Low     14      5    0
    Medium  10     48   16
    High     0      3   22

Overall Statistics

               Accuracy : 0.7119
                 95% CI : (0.6213, 0.7915)
    No Information Rate : 0.4746
    P-Value [Acc > NIR] : 1.502e-07

                  Kappa : 0.5209
```

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  100

Number of Support Vectors:  341

Confusion Matrix and Statistics

          Reference
Prediction Low Medium High
    Low     20      7    0
    Medium   4     45    8
    High     0      4   30

Overall Statistics

               Accuracy : 0.8051
                 95% CI : (0.722, 0.8722)
    No Information Rate : 0.4746
    P-Value [Acc > NIR] : 1.476e-13

                  Kappa : 0.6913
```

Even though all 3 models achieve over 70% accuracy, the linear and polynomial kernels have more difficulty separating countries of medium and high happiness levels, especially for those countries that fall in the group of high GDP and high healthy life expectancy. On the other hand, the model using the radial basis kernel is able to segregate different levels of happiness in this group. Using a higher gamma value, this model is able to capture the complexity and shape of the happiness data. In addition, given that this model also uses a higher C parameter, the problem of overfitting is not a big concern. As a result, the SVM model with the radial kernel can be considered the best performing one out of the three. Since this model only uses 2 predictor variables, thereby confirming that GDP and healthy life expectancy are 2 factors that highly impact a country's happiness level.

Supervised Learning Methods Accuracy Comparison

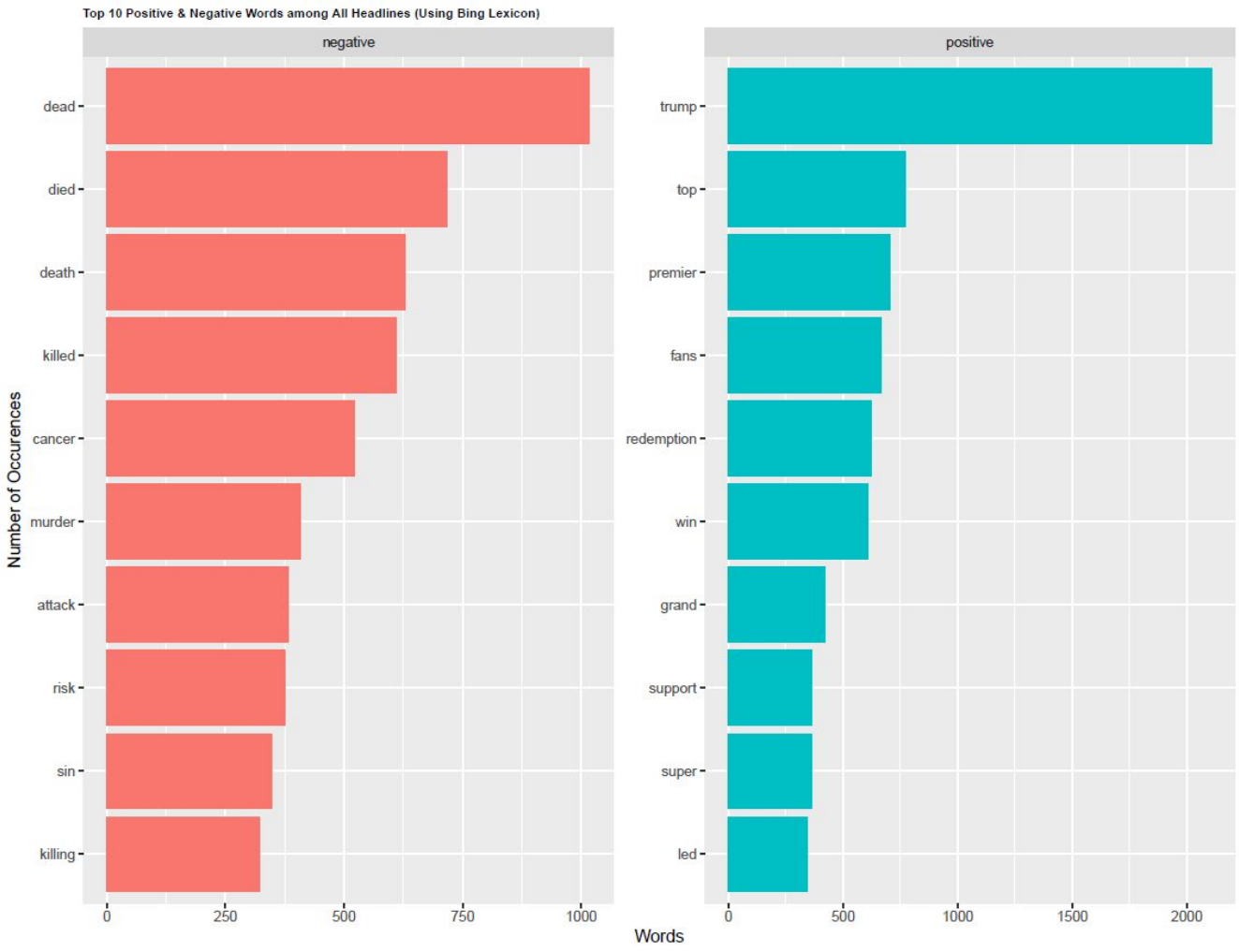
Details and accuracy for each supervised methods used in this analysis are summarized in the table below.

Classification Model	Predictors	Prediction Classes	Training Data	Testing Data	Accuracy %
Naïve Bayes	GDP, HLE, Social Support, Freedom of Choices, Generosity, Corruption Perception	Happiness Level (Unhappy, Happy)	2013 to 2017	2018	81%
Naïve Bayes	GDP, HLE, Social Support, Freedom of Choices, Generosity, Corruption Perception	Happiness Level (Low, Medium, High)	2013 to 2017	2018	75%
Decision Tree	GDP, HLE, Social Support, Freedom of Choices, Generosity, Corruption Perception	Happiness Level (Low, Medium, High)	2013 to 2017	2018	72.03%
SVM (Linear)	GDP, HLE, Social Support, Freedom of Choices, Generosity, Corruption Perception	Happiness Level (Low, Medium, High)	2013 to 2017	2018	73.73%
SVM (Polynomial)	GDP, HLE, Social Support, Freedom of Choices, Generosity, Corruption Perception	Happiness Level (Low, Medium, High)	2013 to 2017	2018	71.19%
SVM (Radial)	GDP, HLE	Happiness Level (Low, Medium, High)	2013 to 2017	2018	80.51%

Text Mining Analysis

With sentiment analysis, using the Bing and NRC lexicons, dissecting over 50,000 headlines reveals insight into how positive and negative overall, and how emotionally charged, these headlines seem to be, and what emotions appear most often. Using both the Bing and NRC lexicon in R, we explore the headlines both across all of the countries and in Africa, and we compare the happiest country, New Zealand, with the least happiest region in the dataset, Africa.

The Bing lexicon calls attention to the various sentiments that are present in all headlines - in the visualization below, the top ten most positive and most negative words are shown:

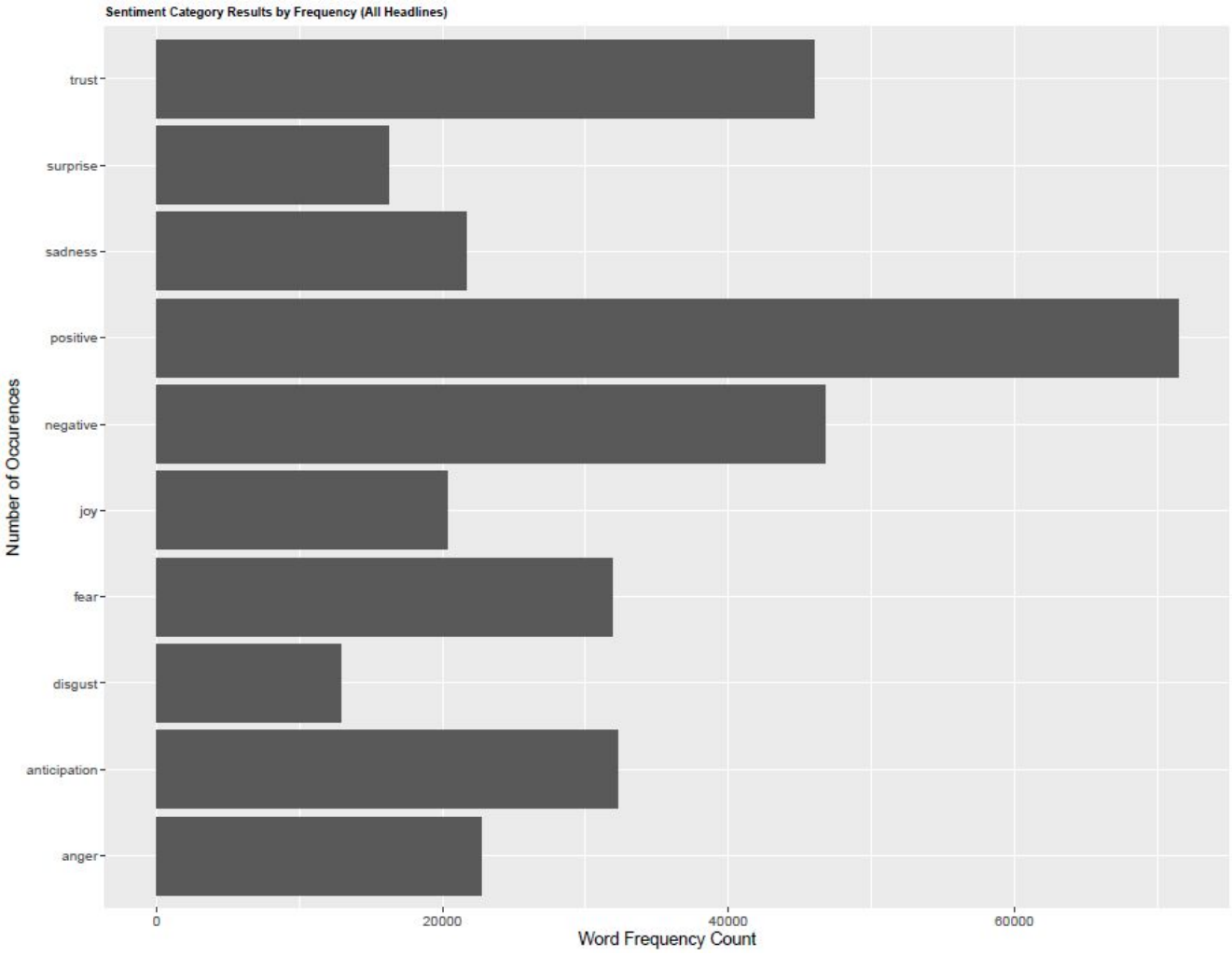


It is highly interesting here that, innately, the Bing lexicon considers “trump” to be a positive word. Additionally, it is most unfortunate and perhaps even distressing to see here that the most commonly used negative words all refer to death, dying, diseases, and violence. The most positive words across all headlines seem to refer to, most likely, soccer - perhaps during this month of headlines, important soccer games or tournaments were occurring; in the United States we commonly

forget that soccer is so popular around the world, bringing countless people much in the way of happiness. Future analysis could delve into either removing “trump” or reclassifying it, depending on the nature and goals of the analysis; or, look at just “trump” entirely, as that seems to have quite a lot of potential for commentary on world affairs.

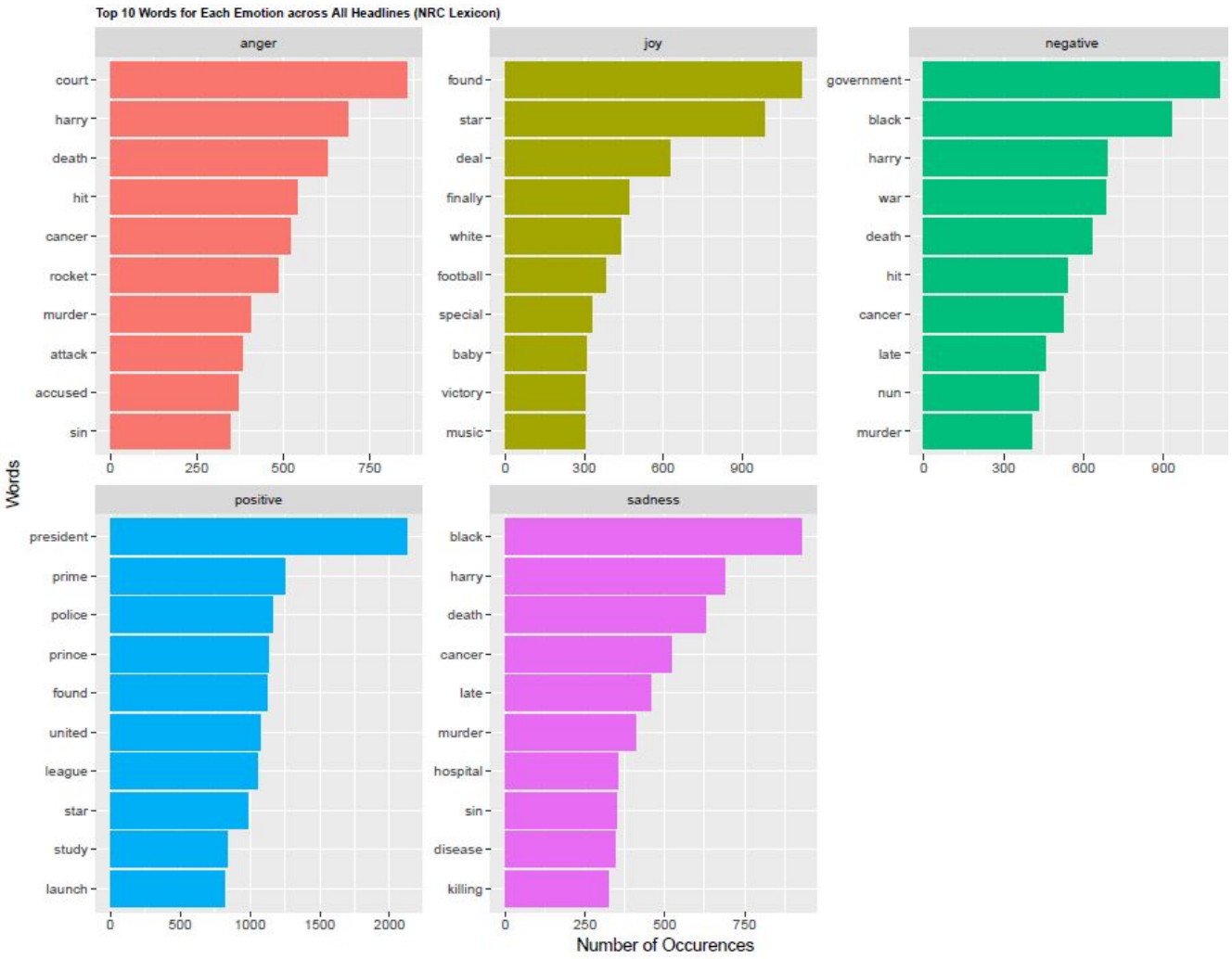
Continuing sentiment analysis, next we turn to the NRC lexicon. Initial NRC analysis finds that there seem to be, overall, more negative words than positive ones -- a finding that is consistent with studies’ findings and the opinion of Steven Pinker, world-renowned cognitive psychologist and author. Writing in 2018 for the Guardian, in an article called “*The media exaggerates negative news. This distortion has consequences*,” Mr. Pinker writes:

> *“Media scholars who tally news stories of different kinds, or present editors with a menu of possible stories and see which they pick and how they display them, have confirmed that the gatekeepers prefer negative to positive coverage, holding the events constant.” (Pinker, 2018)*



Above: A bar graph showing the results of NRC sentiment analysis of all headlines' sentiments among all countries and the region of Africa.

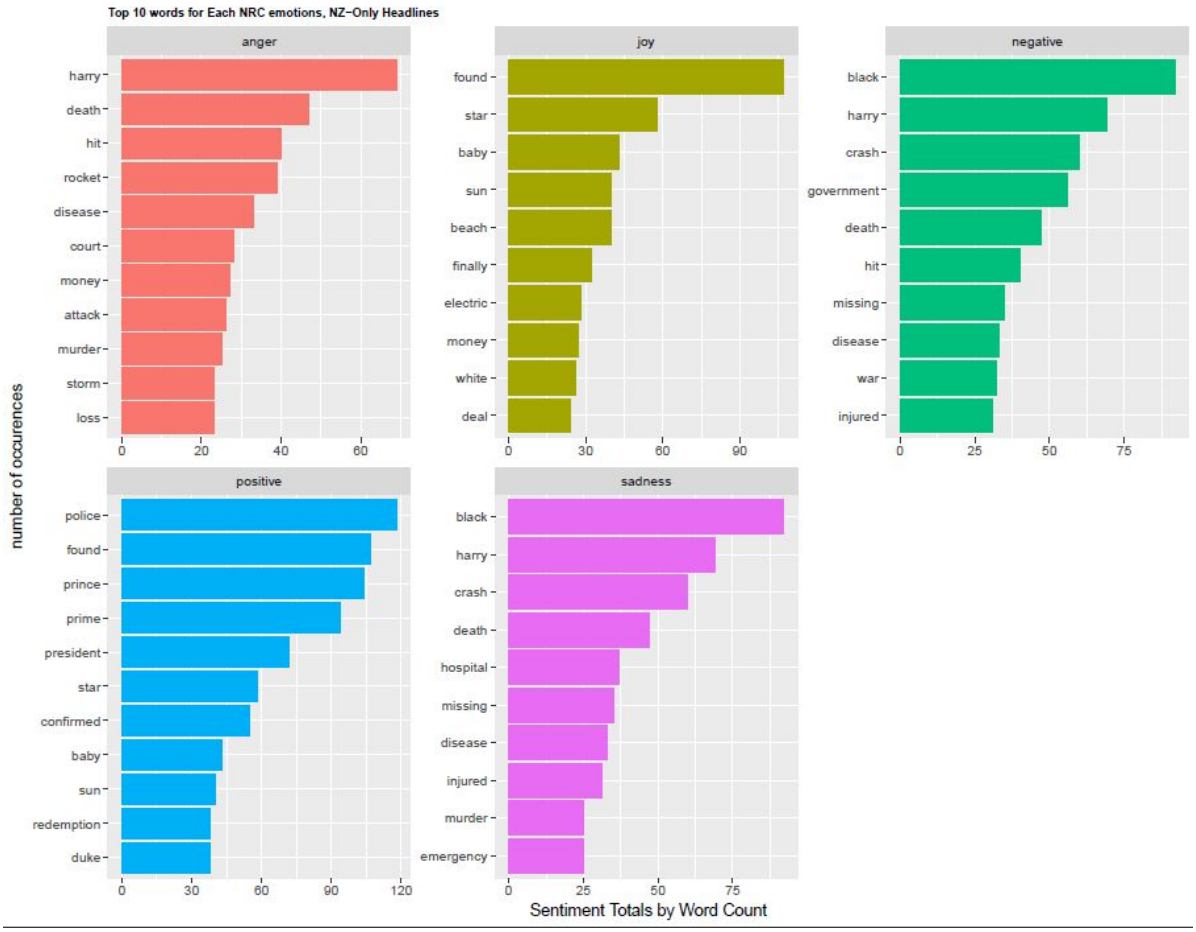
Delving deeper into NRC sentiment analysis, we use ggplot and dplyr to identify the top 10 words that contribute the most to which sentiments, in all headlines aggregated:



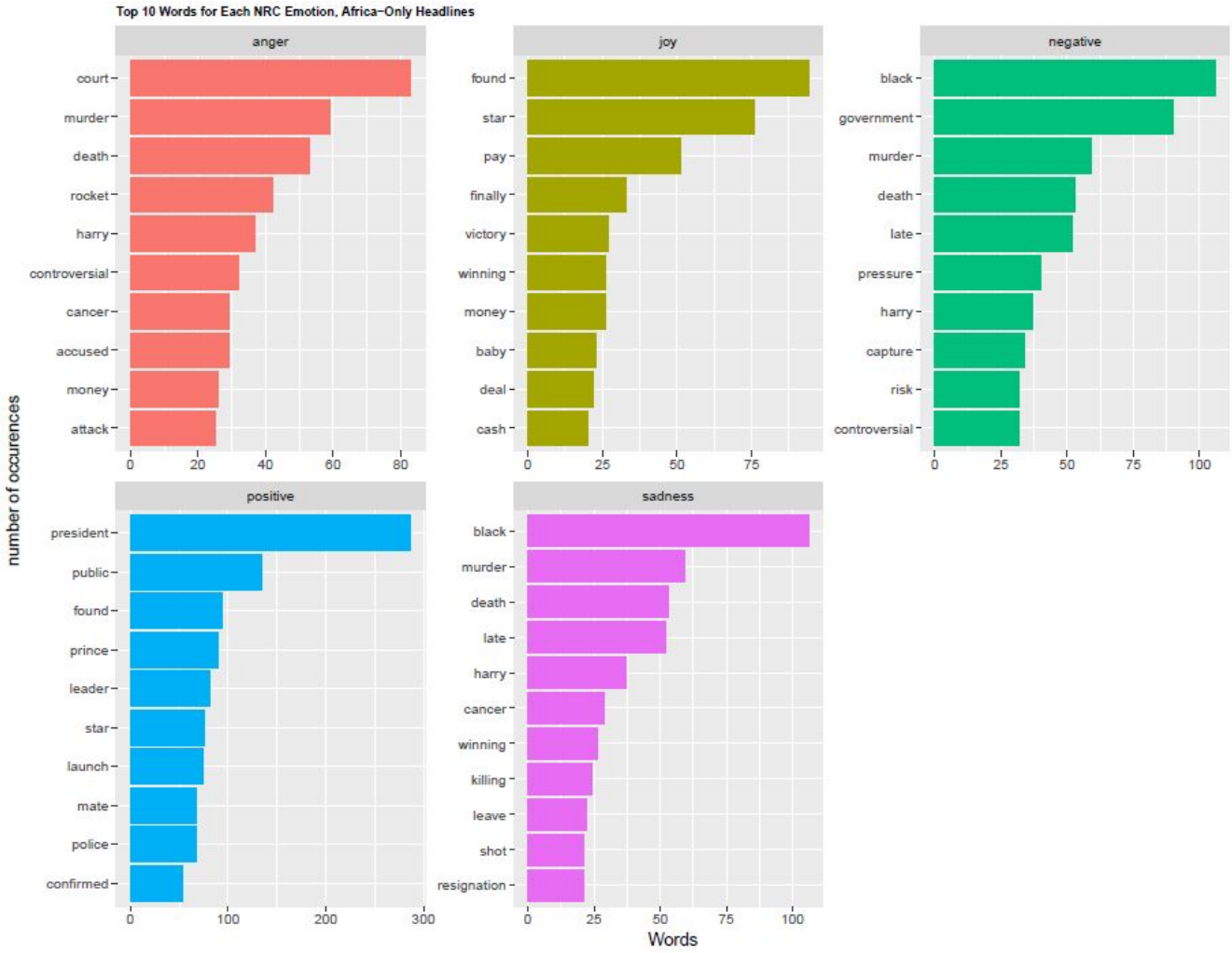
Above: Bar charts displaying the results of identifying the top 10 words contributing to each sentiment in the NRC lexicon, among all headlines in the dataset.

Here, results show that words pertaining to death, dying, violence, and disease dominate the negative sentiments in world headlines. Interestingly, and perhaps unexpectedly, the NRC lexicon defaults to the word "government" being a negative word, yet "president", "prince, and "police" as a positive ones. This finding could merit further exploration and adjustment. Perhaps here more so than any other area, the data and analysis techniques would require subject matter expertise, and careful attention to translation and context. The word "black" for example is considered a top sad word; this is arguably inaccurate, yet could be entirely true. Individual analysis of headlines to confirm or refute this emotion would be time-consuming but invaluable to this sort of technique.

To further explore this dataset, especially in relation to the World Happiness Report, we examine the NRC sentiment analysis results in greater detail by applying it to the happiest country in the dataset - New Zealand - and comparing its sentiment to the least happy region, Africa. The charts below display the results of NRC analysis on New-Zealand versus Africa's headlines:



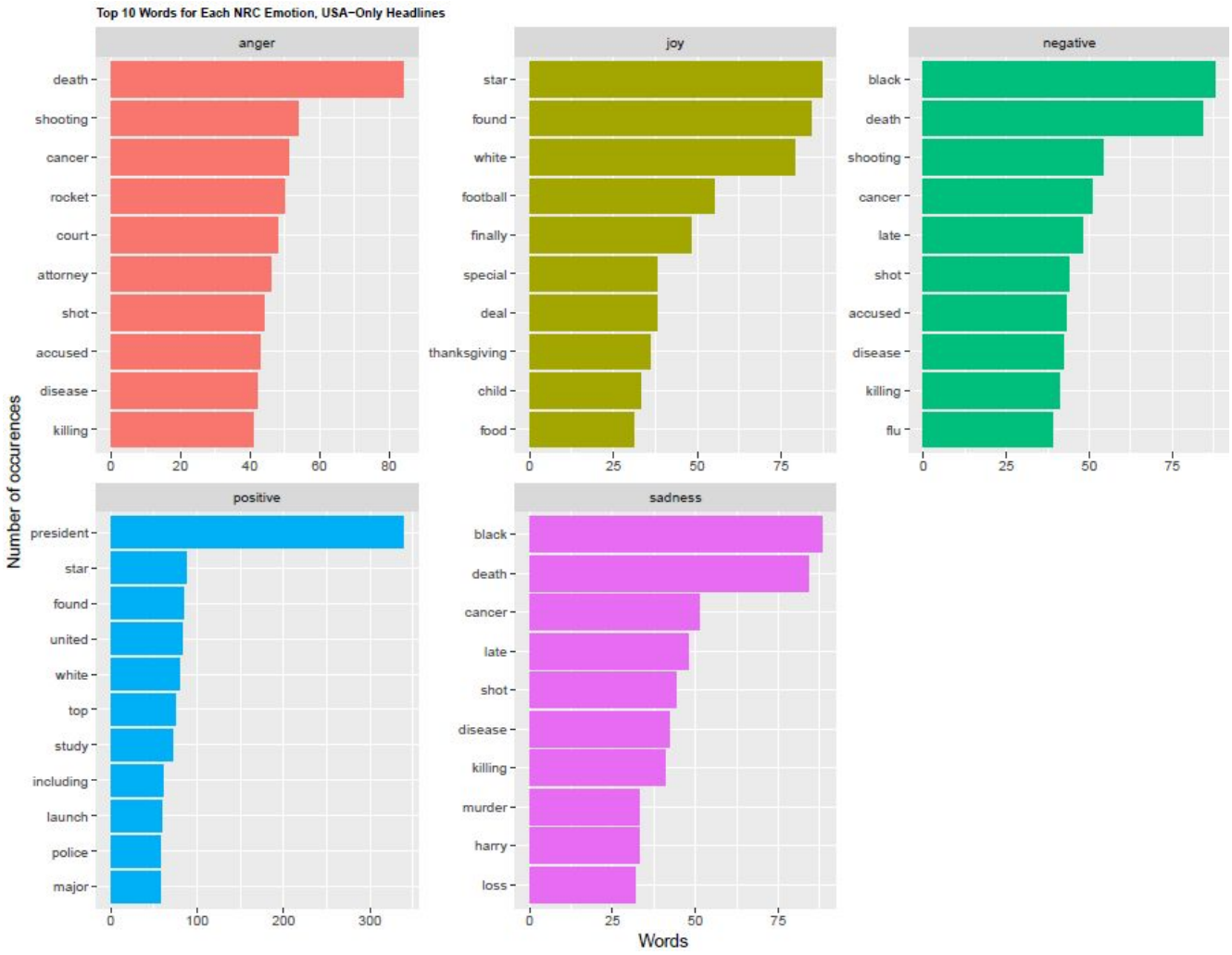
Above: The top 10 words contributing to each sentiment in the NRC analysis of New Zealand's headlines only.



Above: The NRC sentiment analysis results for Africa-only headlines, for the top 10 words contributing to each NRC sentiment.

This analysis shows that, intiutively, it appears that overall, New Zealand's NRC sentiment has more positive words. Additionally, Africa's top sentiment words especially for negative emotion and sadness all contain more references to violence, including "murder" and "death" for example, occuring more commonly than New Zealand's. As a point of interest as well, the NRC lexicon considers the word "harry" to be an angry sentiment word. This word in world headlines seems to most often refer to Prince Harry of England. Future analysis should prune and fine-tune these words for a more in-depth understanding of context and true meaning behind and among headlines.

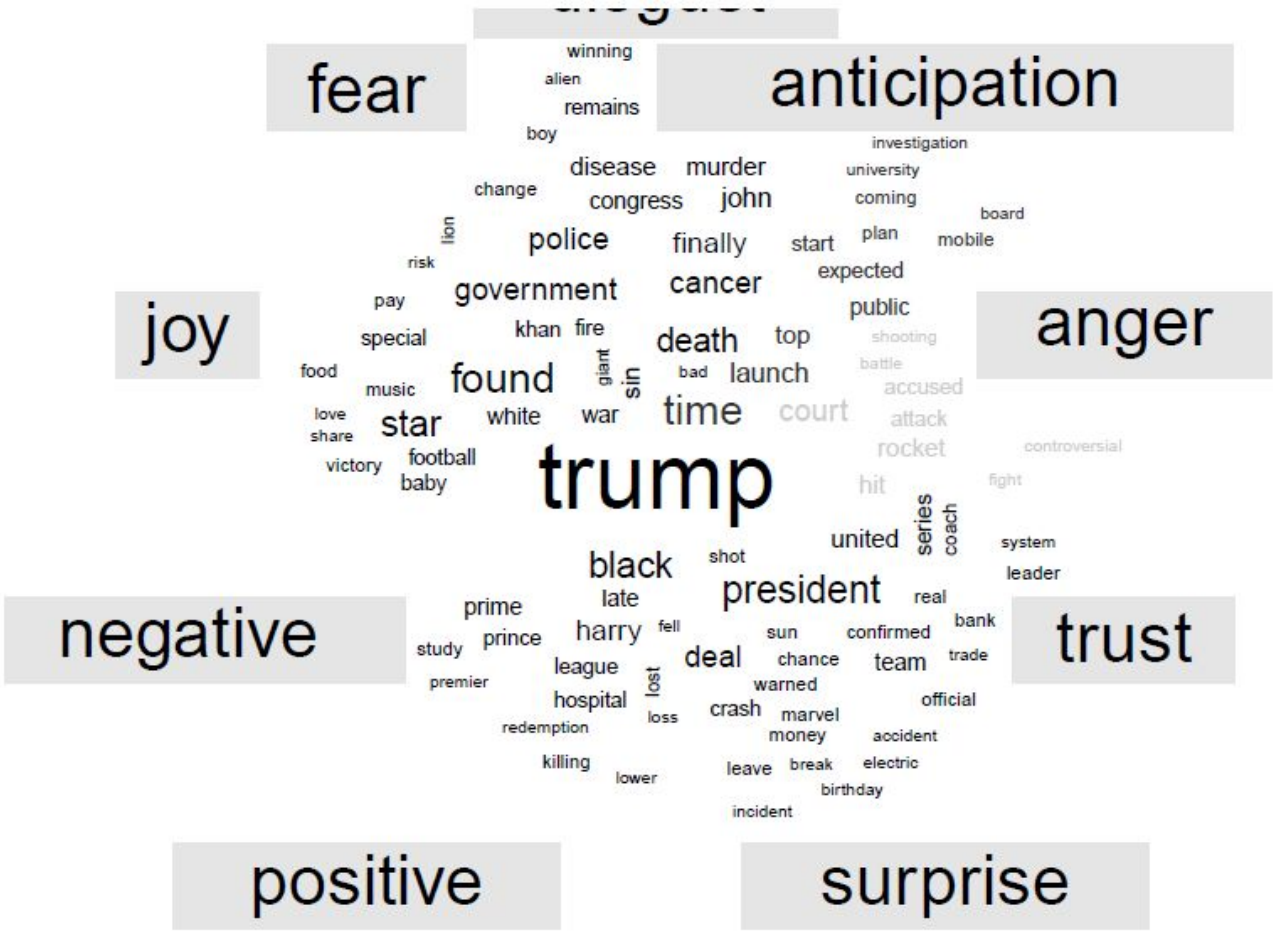
Next, NRC sentiment analysis is applied to the headlines for the United States, as the authors are all United States citizens residing here in the USA - the chart below contains our findings:



Above: NRC sentiment analysis results for United States-only headlines.

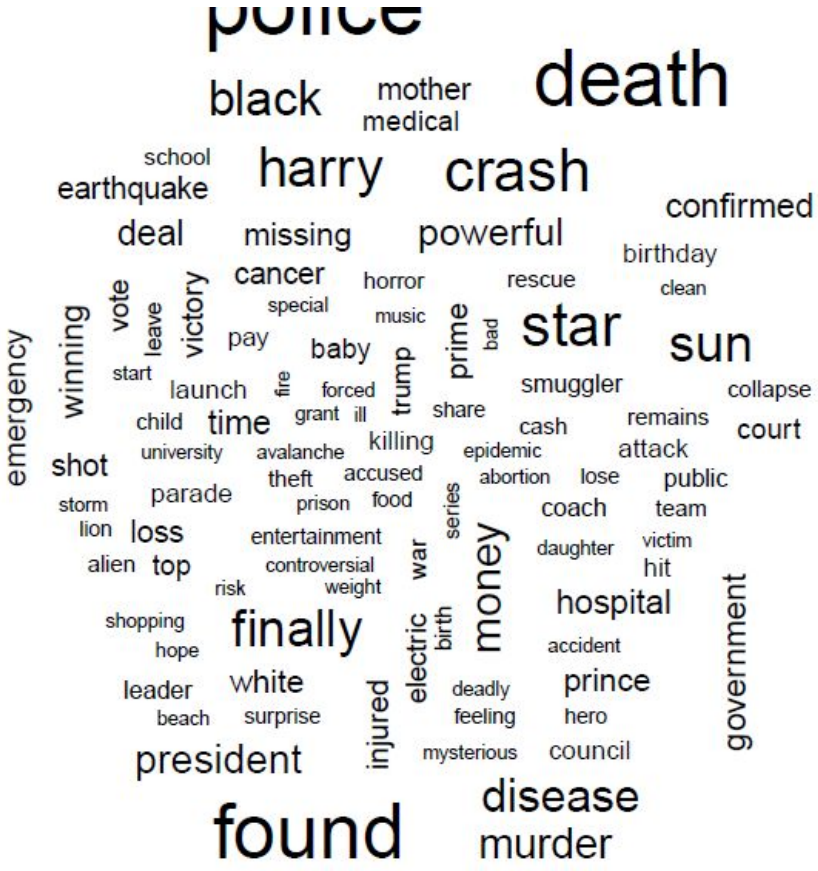
One finding that stands out for the United States' results is, unfortunately, the prevalence of the word "shooting" and "shot" across Sadness, Anger, and Negative sentiments. This seems to be the highest placement for "shooting" among all countries, and even higher than Africa. This finding alone is worthy of in-depth examination, and may be a commentary on not only how all-too-common shootings are in the United States, but how much they are covered in the USA's news media. Another interesting finding is that the NRC lexicon considers "white" to be a Joy sentiment word, and "black" to be both Negative and Sad. Because of the many nuances and varying definitions these words connote and create, in order to fully understand the context, these words should be carefully examined, given the implications for example for racial commentary inadvertently in these results -- future analysis should break down "white" and "Black" to a much deeper, if not even granular, level. One could accomplish this perhaps with bigram and trigram analysis across many USA headlines over a longer period of time than one month.

Finally, to gain a visual understanding of our text mining results, a word cloud is generated for our happiest country, least happy region, and the United States, as a point of comparison. Below is a word cloud in which all NRC sentiment results for all headlines and all words are displayed:



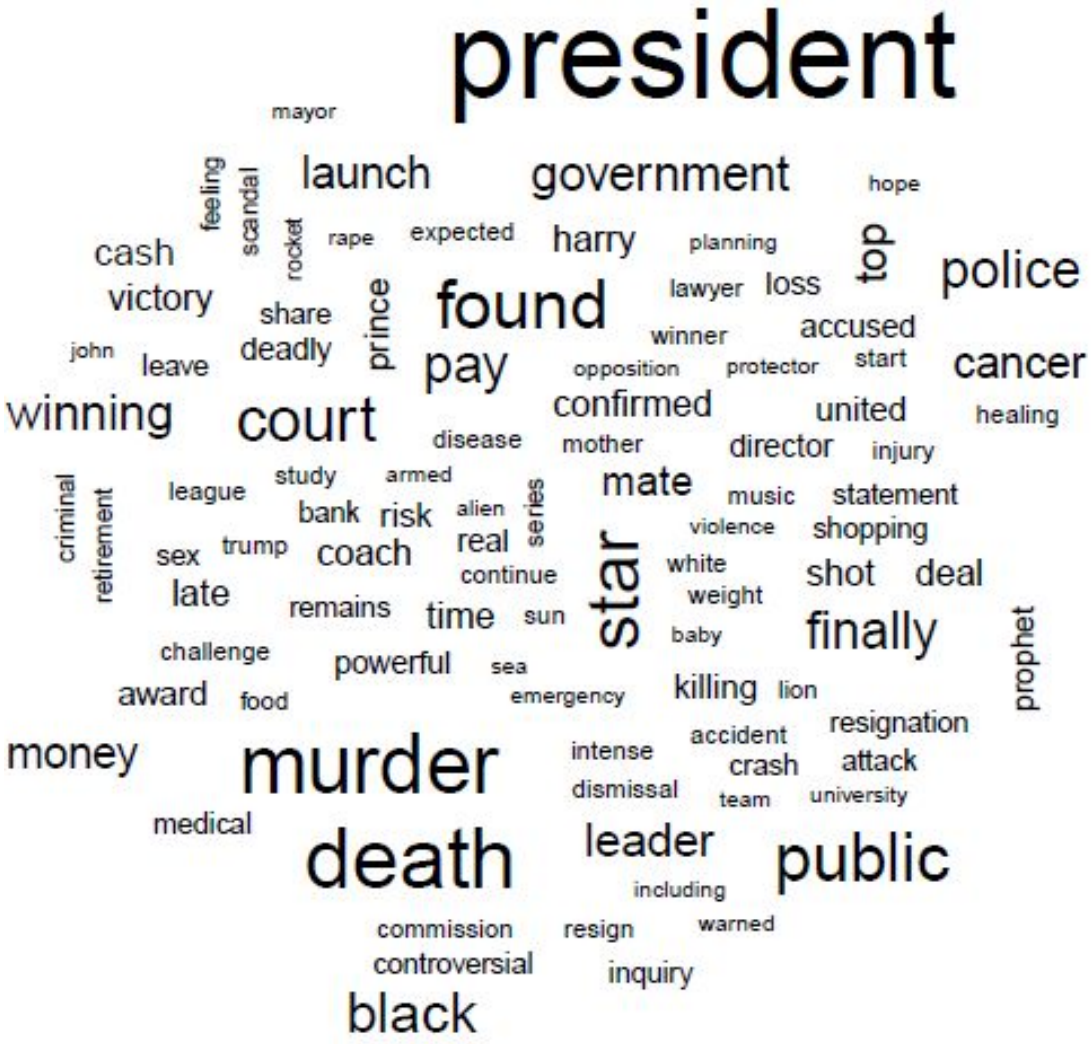
Fascinatingly the word “trump” is central to the word cloud, indicating its prominence as the top-most used word. The shape of the word cloud further supports the observation that more negative words appear across headlines versus positive ones.

To compare this to New Zealand, we generate word clouds for the happiest country in the headlines dataset, and Africa, the least happy region, and the United States. Below is a word cloud for New Zealand's headlines:



Above: New Zealand's word cloud, displaying the top 100 most frequently used words in New Zealand-only news headlines.

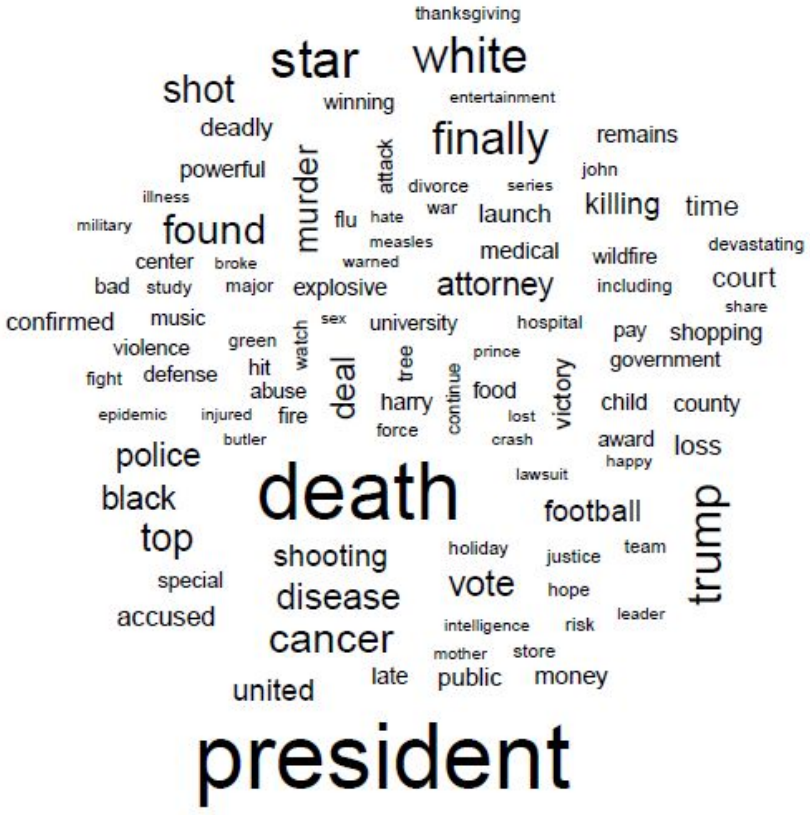
New Zealand's word cloud on the whole contains more positive words than Africa's and the United States, which is an interesting result -- Africa's word cloud follows:



Above: Africa's word cloud, showing its top 100 most commonly occurring words.

In Africa's word cloud, more words refer sadly to violence, death, and corruption; this seems to mirror events occurring across Africa's many countries and regions, where corruption, violence, virulent disease, and natural disasters are all too common occurrences and realities of its citizens' lives. The presence of the word "rape" is a strong point of comparison to New Zealand in particular. The many words referring to violence in particular seem to relate to Africa's life expectancy being so low versus many other regions in the world, and support the World Happiness Report's findings related to life expectancy versus happiness.

Finally, we compare New Zealand and Africa's word clouds with the United States:



Above: The word cloud results for the top 100 words for the United States-only headlines.

Unlike Africa and New Zealand, among the top 100 words for the United States the word "shooting" appears. This supports our earlier finding that "shooting" occurs commonly in United States headlines. The word "death" appears almost as frequently as the word "president" in this word cloud (indicated by the size of the words in comparison to one another). Many of the words in the US's word clouds relate to the president and political happenings, an indication of the extent to which political events dominate headlines in the USA.

CONCLUSION

On the basis of key indicators alone, such as GDP, Social Support, Healthy Life Expectancy, Freedom of Choice, and Generosity, distinct regional groups and patterns emerge that in turn can be mapped to general happiness levels within countries. From observing these patterns, it is apparent that accurate predictions can be made with regards to how happy citizens in any given country are likely to be, provided full disclosure is provided regarding each country's social, cultural and economic factors.

There seems to be a distinct correlation between Healthy Life Expectancy and perceived levels of happiness. Countries that are predicted to be in the least happy category all demonstrate a life expectancy that does not exceed 60 years in age, with most falling below this (the mean life expectancy within this group being 55 years). Moderately happy regions contain some countries that fall below 60, but in general most citizens can expect to live just a smidgen above 60 years of age. In contrast, those countries designated to be the "happiest" enjoy on average a healthy life expectancy at birth that is 15 years longer than those that are least happy, with 70 or more years being the longevity that is expected at birth.

When comparing GDP and Social Support within three prediction categories (low, medium, high), citizens living in countries that have a lower GDP but that have access to higher levels of Social Support appear to experience greater levels of overall happiness.

Text mining of world events via examination of news headlines across 21 different countries overall supports the findings of the World Happiness Report: countries and regions that are lower in Happiness Scores in the report tend to use negative words more frequently. This is a key finding that is supported by previous studies and research, and it indicates that events occurring in each country seem to correlate strongly with the happiness of its citizens. This finding certainly is worthy of further exploration: is that negative words occur because citizens are unhappy, or, is it that the media focuses more on negative events disproportionately, so a country's citizens are constantly exposed to more negativity, creating a feedback cycle of negativity? Another finding these results reveal is that the events happening in the United States have a much larger impact than just on the United States' citizens: our presidential politics affect the sentiment and happiness of millions of people across the world.

One could delve much deeper into the realm of text mining in relation to world happiness - examining social media and headlines across a longer period of time, exploring bigrams and trigrams in relation to one another and sentiment among more headlines, in order to gain deeper insight into the realities shaping the happiness of world citizens.

Through the use of correlation analysis and supervised learning models, GDP and Healthy Life Expectancy are two factors that surface as having the highest correlation to happiness scores.

However, life satisfaction and happiness still vary widely both within and among countries. Instead of drawing associations between survey measures and other supporting factors alone, happiness data should be analyzed over time within the context of trends, and timings of significant events such as financial crises, political elections, or war; all of which may provide even greater insights into the variation of world happiness scores.