## **World Happiness Data Analysis**

Prepared by:

Niat Tesfai

Zach Eshete

Hamza Ali

Marcus Walker

## **Executive Summary**

Our team chose the topic because we believe that the purpose is to survey the scientific underpinnings of measuring and understanding subjective well-being. Our team decided to work on the first World Happiness report that supports the UN high level meeting on happiness and human well-being. The World Happiness is published annually to match with the United Nation’s International Day of Happiness. The World Happiness Report 2016 ranks 156 countries by their happiness levels and this data is important because our wellbeing is impacted by having a sense of purpose and feeling what you do in life is worthwhile. Positive emotions matter more to us than the absence of negative emotion although both are important.

Our team goal is find out what major factors caused country ranks or scores to change between the 2015 and 2016 reports and why countries are changing every year.  We also want to know which countries or regions rank the highest in overall happiness and each of the six factors contributing to happiness and did any country experience a significant increase or decrease in happiness.  We look more closely at our target variable which is happiness score and this data will allow us to analyze it from thousands of individuals around the world and investigate the ways that makes people well-being. We look at the rating and explore the extent to which people have negative or positive affective states in their day to day lives. The most informative attributes from this data are healthy life expectancy, family, freedom to choose what you do in life, generosity and how much people donate to charity, trust (Government Corruption), economy (GDP) In short, the current data could be fun and challenging to work with, but our team is going to find the correlation between each variable.

**Table of contents**

Teams correction…………………………………………………………………………………..1

Project Business or Policy Understanding…………………………………………………...........1

Data Understanding……………………………………………………………………………….3

Data Preparation …………………………………………………………………………………..4

Descriptive Analytics …………………………………………………………………………......5

Expanded Modeling and Predictive Analytics…………………………………………………...15

Enhancing Models and Prescriptive Analytics……………………………………..........15

Evaluation……………………………………………………………………………………......18

Comparison with published results ……………………………………………………………...19

Deployment………………………………………………………………………………………20

**Teams correction from Projects#1 & #2**

Our group made correction on Descriptive analytics, Modeling and predictive analytics and Comparison with documented results. For the descriptive analytics, we used multiple charts as the Instructions say and use different charts to strengthen your conclusions as shown in page **5**. We also used bar charts to support our report and we used the scatter plot that shows R-squared. The graphs and charts we used fits with our data and story. We also made correction on the comparison with the documented results which is onpage 19. We took the feedback that was giving to us and make correction to the modeling and predictive analytics. We address why and how the model the problem. We used simple regression model, gain information and association models to make a stronger relationship with our variables.

**Project Business or Policy Understanding**

Identify, define, and motivate the business or policy problem that you are addressing. What precisely is the data analytics problem? How (precisely) will the data analytics solution address the business or policy problem? What is the use scenario? What features would be useful? How exactly would it add business value?

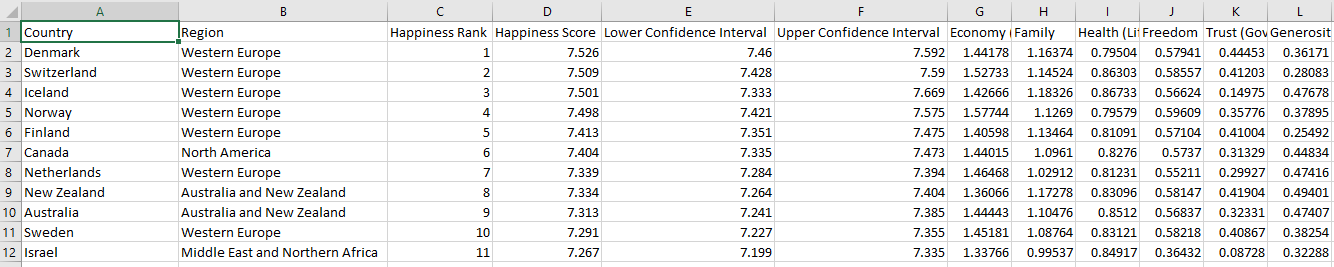
There are various business problems we had to consider when conducting our data analysis. Overall, we had a couple questions and goals we wanted to address with our data analysis. First, we wanted to know what were the major factors of overall happiness? Second, we wanted to know what countries ranked highest in overall happiness. Third, we wanted to know what caused significant changes in happiness between 2015 and 2016 within countries? Our analysis adds value to the globalization trend in business and companies looking to go global or who have already gone global.

When conducting our analysis, we wanted to narrow down which key variables contributed to overall happiness. We found this and discovered that if these key variables had a significant difference between 2015 and 2016, this would cause a significant change in happiness rank. We discovered our two key variables to be Health (Life Expectancy) and Economy (GDP per capita). These variables contributed the most to a country’s overall happiness and had a positive correlation. We also noticed that Health and Economy had a positive correlation between each other with an R Squared of .7. As discussed below in descriptive analytics, we found significant increases and decreases in happiness from 2015 to 2016. This includes Bosnia which increased its Economy and Health levels and saw a significant increase in happiness. Also discussed is Yemen, which has seen to have a decrease in happiness due to the two factors decreasing as well.

So, based on our analysis, we found the most influential variables that contribute to happiness score to be Health and Economy. If there is a significant change in these variables within a country, the effects of this will be felt in the society. We proved this when analyzing significant changes in countries happiness score.

Also, this information is valuable to business because it shares which countries are best to do business in, with, or locate. As the world become a more global market, we can use our analysis as a tool to add value to businesses.

**Data Understanding**



Given the World Happiness Report 2015 and 2016, we have various business problems as well questions to answer.  To understand our data, we must do a simple analysis of our data sets. We are given two datasets which will help us solve our business problem and answer our questions. Our two datasets contain happiness scores and rankings which uses data from the Gallup World Poll. Scores in this data set are based on main life evaluations questions asked in the poll. Questions are based on the Cantril ladder, which asks respondents to think of a ladder with the best possible life. This asks respondents to rate their life from zero to ten. Ten is the best life possible and zero is the worst. These scores were collected from 2013-2016 which makes this data very relevant.

Columns in this data set include country, region, happiness ranking, happiness score, lower confidence interval, upper confidence interval, Economy(GDP), Family, Health, Freedom, Trust, Generosity Dystopia Residual.

Columns following the happiness score take into consideration six factors which make life evaluations higher in each country. These six variables are economic production, social support, life expectancy, freedom, absence of corruption, and generosity. Dystopia doesn't affect a country's happiness score it has values which is equal to the world's lowest national average for each of the six factors. It has no impact it is just used as a reference to compare with low averages. To understand our dataset, we need to understand our variables. Our datasets contain 13 variables and 157 instances. Our first variable is country which is nominal variable. It contains 156 countries. Our second variable is Region which is also nominal. This breaks the data up into regions according to continent. Our third variable is happiness rank which is numeric and ranks each country's happiness from 1-156 based on the world happiness score. The world happiness score variable is numeric and is based on a happiness scale from one to ten. Our lower and upper confidence levels are used to predict means these are both numeric. We also are given a GDP per capita numeric variable which gives us insight into how well a country is doing economically. Family is a numeric variable which gives us insight in how much a country values family. Health care is a numeric variable which is based on access, and quality of health care. Freedom and trust are a numeric variable. This is based on how free do people feel in the given country and how much do they trust the government, Generosity is a numeric variable based on how generous the people are in the given country. Dystopia Residual is our last variable which is numeric an imaginary country that has the world's least happy people. It is imaginary country which has the world’s least-happy people. This variable is used to compare with other countries and evaluate their level of happiness.

**Data Preparation**

**Data Cleaning:**

To prepare this dataset we began by uploading our files onto Weka. We first wanted to understand how much of our data was missing. We went through both files and determined that we had no missing values. Next, we compared the two files we were given to ensure that they data was consistent. Looking deeper into the dataset we found that there was little editing required as the data provided was legible, complete, consistent, and correct. This dataset was very clean.

**Data Integration & Reduction:**

We continued preparing the dataset by finding some of the more informative variables. We ran a few tests to find the most informative variables and compared them to see if there were any relationship among the more informative variables. Next, we looked to see if there were any

variables that we could delete or combine. We then looked at how we could potentially integrate some different variables to form more accurate and perhaps informative variables.  However, we found that we could not delete or integrate variables because of how well classified each variable was. We also were debating on the class variable. Some group members thought it should be happiness rank while others thought it should be happiness score. We decided happiness score was the best class variable and would be the clearest cut choice.

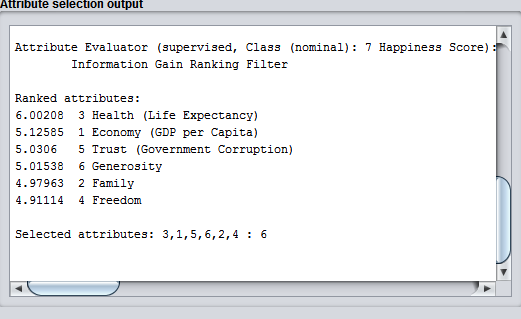
**Data Transformation:**

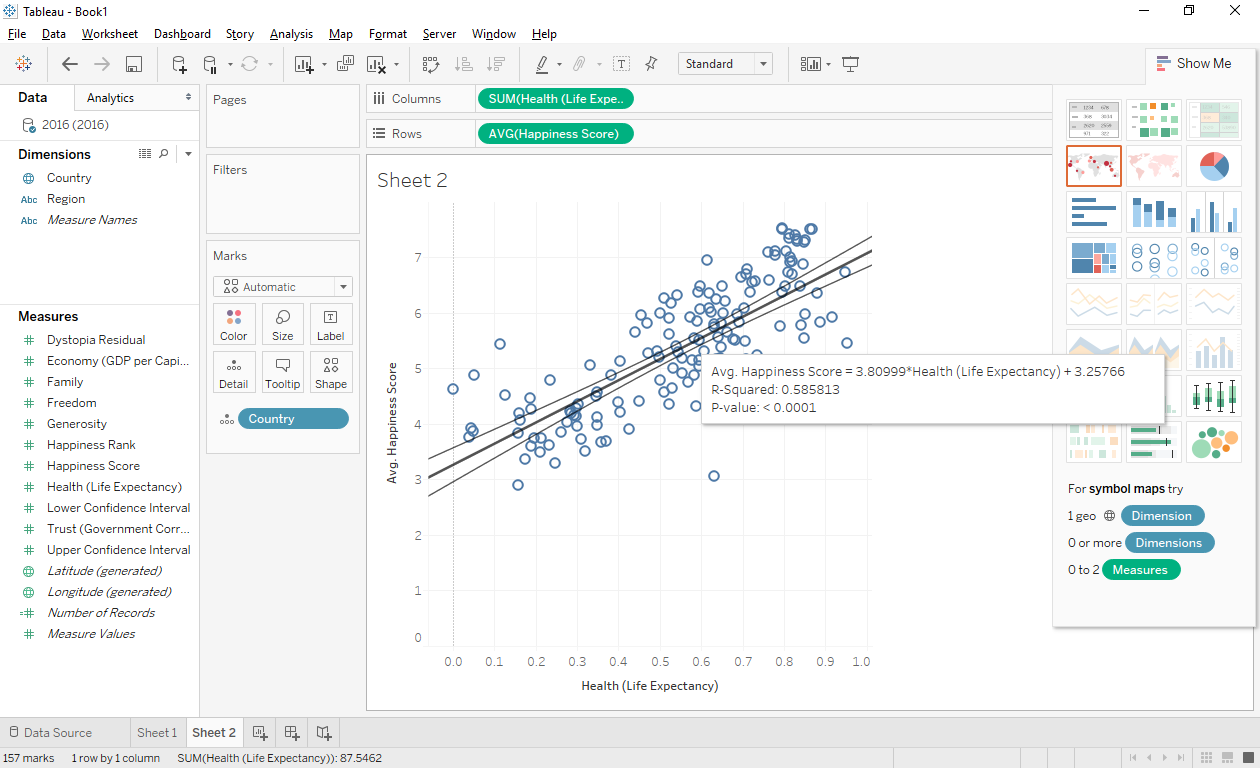
Since our data was already ranged out for each year and was clean, we found it easy to comprehend and allowed us to reduce the time we spent preprocessing. We then started to analyze our data.

**Descriptive Analytics**

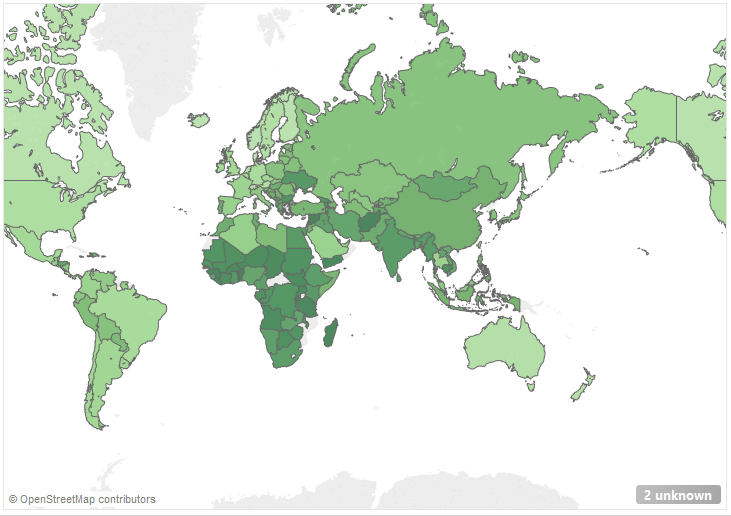
Happiness is a very important aspect of one's life. Using analytics, we can dig dip into factors on what makes a person truly happy. This is where descriptive analytics comes in. Descriptive analytics give is important when describing what has happened with our data. Given our two datasets, the tools we used were Weka and Tableau to visualize our data. These gave us the best findings when exploring our dataset. Weka and Tableau allowed us to gain insight in our data set and understand what has happened in 2015-2016. Simply, we have found 3 important variables that lead to happiness in the world. These variables include health, economy, and government trust.

When using Weka, we changed the class variable to happiness score (numeric).  Then we used the selected attribute evaluator to analyze this data set. We chose to run the information gain evaluation because we were really interested to see which were the key variables which impacted the happiness scores among countries. When running this evaluator, we found a lot of valuable information. The variable with the most information gain (Figure 1) was health (life expectancy). This ranked first and shows us that Health is very important to one's happiness. This makes logical sense. The second most informative variable is economy (GDP per Capita). The third most informative variable we found was trust (government corruption). The fourth most informative variable is generosity. The fifth most informative variable was family. Finally, the last most informative variable is freedom.  From the information gain we are able to see how these results correlate with the overall information gain. We shall focus in the three most informative variables and analyze these. The three most informative variables in order are health (life expectancy), economy (GDP per capita), and trust (government corruption).



**

As we can see, health (life expectancy) is the most informative variable in this data set when running information gain. We then used this information to see if there was a correlation between our target variable and our most informative variable. To determine this, we graphed happiness score and health (life expectancy). From this, you can see a positive correlation. As health (life expectancy) increases so does the overall happiness score. This is very informative and gives us insight into why people are happy in certain countries. Health care plays a major role. However, this isn’t the only factor that has a strong correlation with a country's overall happiness.

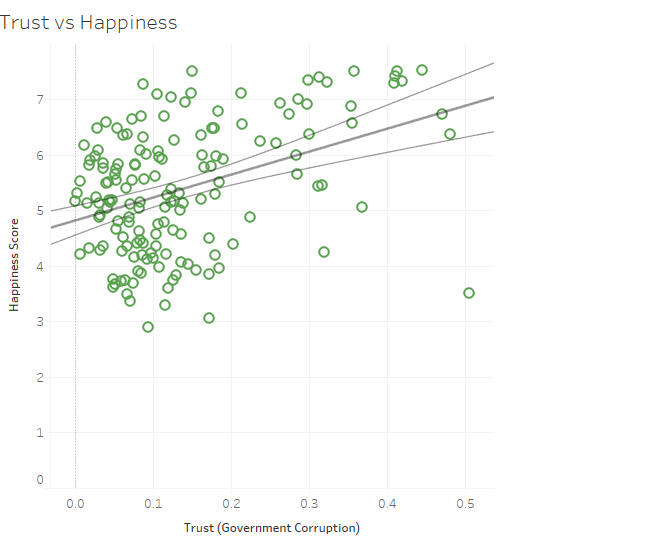
From Weka, we also found that there was a correlation between our target variable (happiness score) and our second most informative variable (economy). We used Tableau to see visualize this correlation. We also found a positive correlation between happiness score and Economy(GDP). As a country's economic score and GDP increases, so does their overall happiness. This is logical due to the fact that if one's GDP is high then their happiness is likely to be high. You will have less worry and doubt. Vice versa, if your GDP per capita is low your happiness is likely to be low too. You will be struggling to get by in life and have lower happiness due to this.

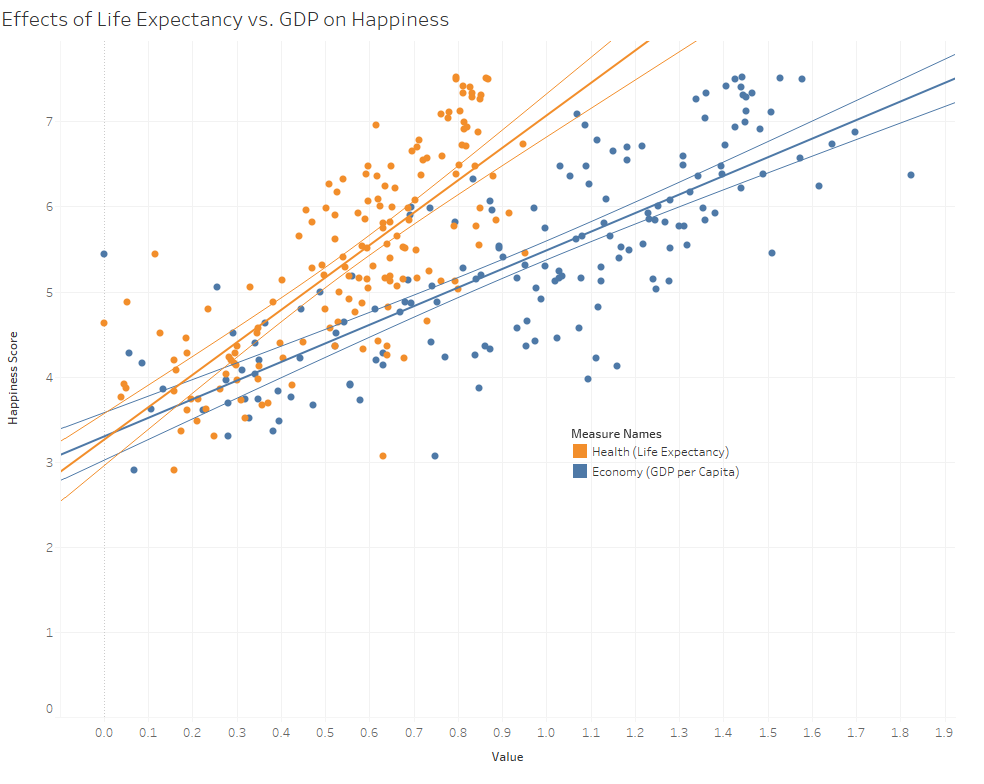
Our third most informative variable was trust (government corruption). The higher this variable, the more that the country trusts its government. We then visualized this information in Tableau and found a trendline to get an accurate portrayal of this graph. After implementing the trend line, we found a positive correlation between trust and happiness. The more the people trust their government, the happier they are. This makes logical sense because if you are upset and have little trust with your government, the less happy you are likely to be. However, this data was collected before Donald Trump was elected so the U. S’s happiness score may not be recent. We also found correlation between increases and decreases from 2015 to 2016.

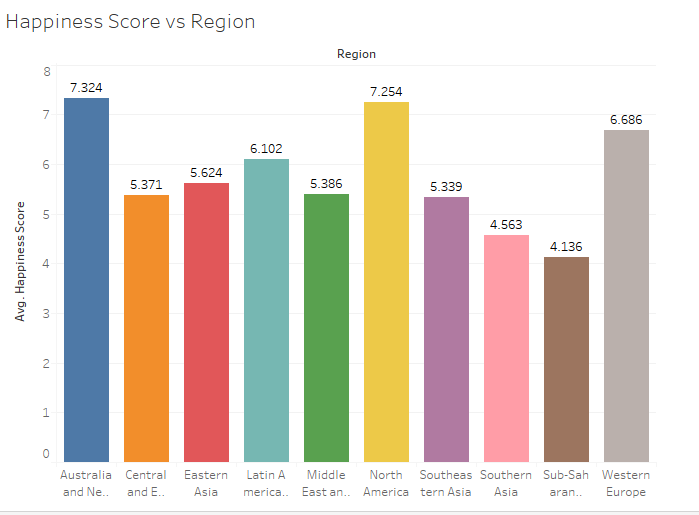
When analyzing the data, we found that some happiness scores decreased, increased, or remained the same from 2015 to 2016. We wanted to analyze what factors made a country happier or sadder. By analyzing the data from 2015 to 2016 we could analyze the major factors that contribute to an increase or decrease in happiness scores. Although there are some outliers, the majority of correlations behaved the same. When a country's happiness score and rank increased, their three most informative variables also increased.

Let's take a look at Bosnia. We wanted to know how did Bosnia increase its happiness rate from 4.949 to 5.163. When looking at its health, economy, and trust scores, all of these increased. These made the people of Bosnia happier. For example, health (life expectancy) increased from .708 to .791. Also, economy (GDP per capita) increased from .832 to .934. Finally, the last most informative variable was trust (government corruption). We see that Bosnia increased this score from 0 to .002. This gives us a description in how Bosnia increased their happiness score. Other countries follow this trend as well, Bosnia is just one of many examples. This shows us how our most informative variables are crucial.

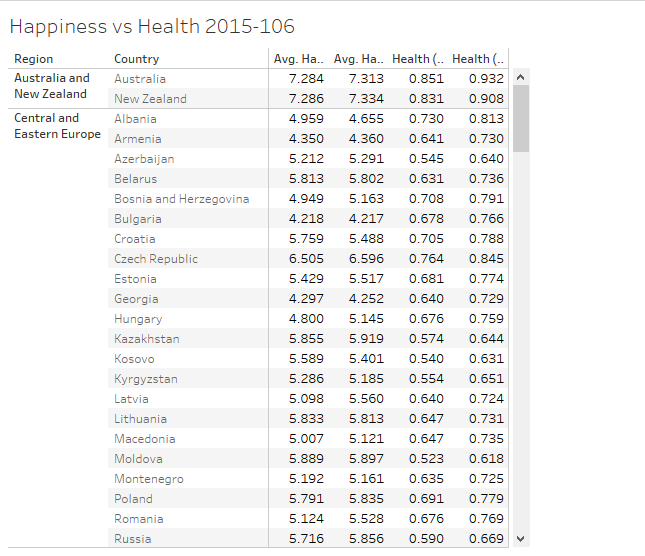
Another thing we can analyze is why countries happiness scores decrease. We analyzed the three most informative variables again. We found a country whose happiness score has decreased. Yemen had a happiness score which decreased from 4.077 to 3.724. The three most informative variables decreased as well. For example, Yemen's average health decreased from .401 to .310. Also, its economy (GDP per capita) also decreased from .546 to .579. Finally, the last informative variable trust decreased from .079 to .059. This gives us a descriptive picture into what key variables have led to countries increase and decrease in happiness scores. We can determine that if one increases its health, economy, and trust in a country, they will live a happier life. On the other spectrum, if your health, economy, and trust in a country decreases, you will be less happy.

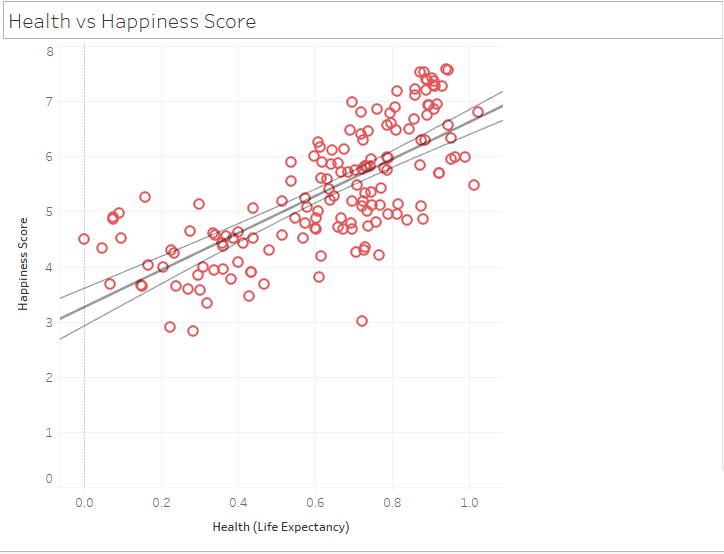


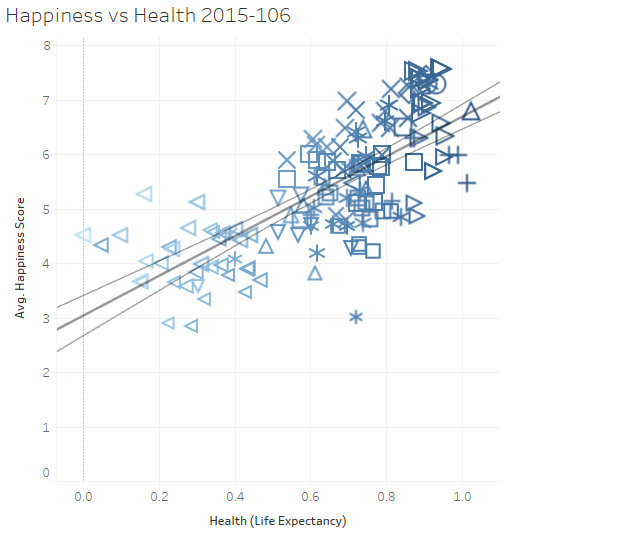


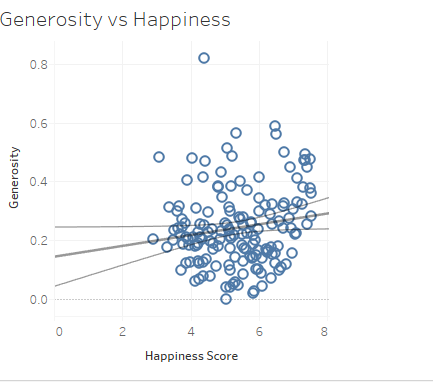
The graph below shows the relationship between our target variable (Happiness score) and region. According to the graph, Australia has the highest average happiness score and North America follows next. All of the top four regions score highly on all the main factors found to support happiness: caring, freedom, generosity, honesty, health, income and good governance.  averages are so close that small changes can re-order the rankings from year to year. 







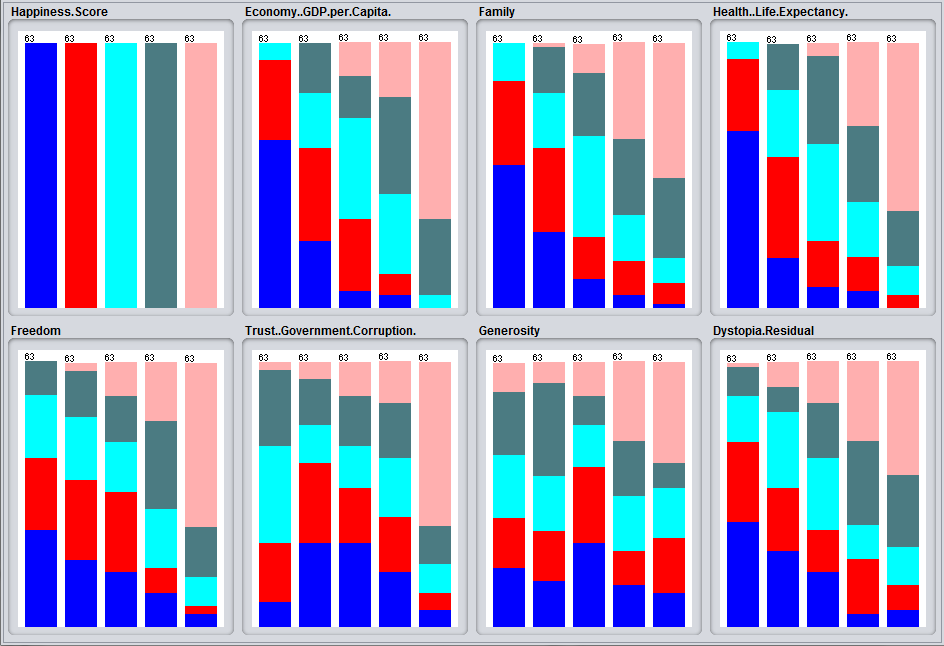




**Expanded Modeling, Predictive Analytics, Enhancing Models and Prescriptive Analytics**

The dataset is constructed in such a way that it would be easy to construct an arbitrarily perfect multiple linear regression from all continuous numerical regression to Happiness Score, based on the way Happiness Score is explicitly constructed from those regressions.  Much like cheating at most things, this would be boring and not useful.

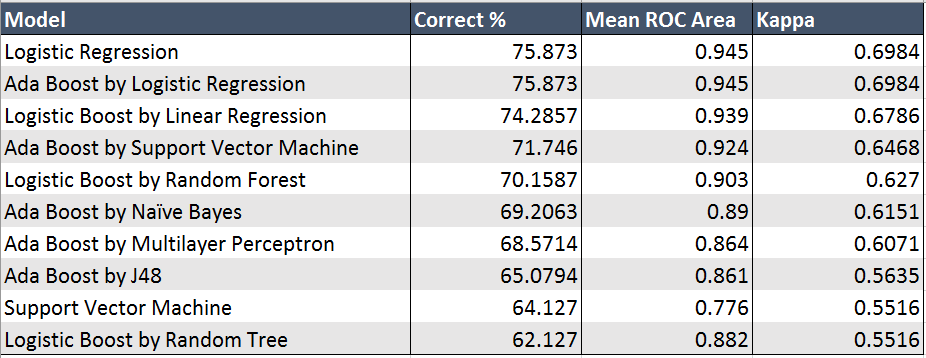
To remedy this problem, we chose to exclude some non-ordinal categorical data for the purpose of predictive modeling, and discretize the remaining fields into five equal frequency bins.  Specifically, the fields Country, Region, and Happiness Rank were excluded.  Happiness Score, Economy (GDP per capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, and Dystopia Residual were discretized into five equal frequency bins.



Field included in model discretized to five equal frequency bins

As a caution, association analysis was performed on the discretized variables, but no rules of significance were supported.  We proceeded under the assumption that no significant useful associations exist among the variables.

This uniformly discretized format lends itself well to logistic regression models and support vector machines.  These models, as well as others, were tested for their ability to accurately predict the quantile of Happiness Score from the quantiles of the seven remaining variables.



Results of modeling methods for predicting quantile of Happiness Score

As shown in the above figure, the highest performing methods are Logistic Regression and Ada Boost by Logistic Regression, which perform identically in terms of accuracy, mean ROC area, and kappa statistic.  Adherence to relative simplicity demands that we consider Logistic Regression to be our best model.

These results also demonstrate the curious phenomenon that, for this dataset and discretized in this way, boosting fails to improve the performance of the models and, in some cases, actually diminishes their effectiveness.  This runs completely counter to our intuition, and we are not entirely sure how to explain it.

The use of this model prescriptively entails some difficulties.  Prescriptions appropriate to market analytics or financial forecasting fail to map to the domain of aggregate happiness at the national level.  None of the knobs turn simply and, even then, not easily.  Neither are these knobs any that a state is not already trying to turn, and so the suggestion to turn them is, in itself, unhelpful.

What our work provides in terms of prescription is a technique for weighting indicators.  An agent interested in predicting when aggregate happiness levels might change could gain greater insight into indicators to watch for that are more accessible to measurement.  Any number of agents may be interested in watching for such indicators: governments, nonprofit NGOs, even businesses interested in operating in places like to become or stay generally happy.

**Evaluation**

Our results differed than our initial thoughts. However, after the different modeling techniques applied, we understood the errors in our initial hypothesis. We were careful to avoid over fitting and other negative influences through our use of information gain ratio and carefully removing and combining certain attributes such as happiness rank to get a clear reading on the relationship amongst our most and least informative variables. We further compared our most and least informative variables with one another to determine the coefficient of determination amongst variables to each. This allowed us to evaluate our results accurately.

Our techniques of trying multiple models with and without boosts allowed us to get a better understanding of how we would be able to evaluate and later implement and deploy our findings. Since this dataset was on an abstract concept, we considered all the different areas of business that our dataset applies to and choose two that would ultimately affect all the others: globalization and government.

We understood that many different companies have been expanding to new areas globally. With globalization comes new opportunities for success and with our evaluation of the dataset we believe that we could help any company expand their business. Our evaluation of the dataset has allowed us to understand the relationship among the most important factors of happiness and how they relate to each other. This understanding of the factors of happiness would potentially allow a company to decide where they would want to relocate open new stores.

We believed that basing where to open a new location on the world happiness report has the opportunity to increase ROI. Through careful analysis of our report, a company could choose where to open a new location based on a country's overall happiness score. For example, if a retail shoe company wanted to open their first international store but couldn’t decide in what country, they would be able to utilize the models we created to determine a new location based on the country's economic standing, population, and health.

**Comparison with published results**

Through this dataset, we have discovered many relationships across happiness and our informative variables. We found that as certain variables increased such as health, economy, and trust of government, overall rank of happiness increased. This was substantial as it gave us an understanding of happiness on a macro level. While manipulating and plotting certain grafts we realized that there certainly was a linear relationship with these informative variables. This can be demonstrated by our analysis of Yemen. Its happiness rank decreased due to the decrease in other factors such as health, economy, and trust of government. This illustrated the relationship mentioned between happiness. However, we noticed that as we got closer to the top rankings and countries become more even across some of the informative variables, some of the smaller variable were the differentiators.

This was an important discovery as it leads us to understanding how the top countries were ranked. We noticed that while some of the attributes such as happiness score, economy, and health were very similar among countries, usually within a tenth of a difference from each other, other countries were leading in some areas substantially, particularly the social areas such as generosity, social support, and freedom.

Our findings were aligned with the findings of the Andrew E Clark in the World Happiness Report 2016. We found that our findings of ranking amongst top leaders being differentiated by social aspects very similar to how the Sustainable Development Solutions Network was able to determine how they classified happiness.

        A difference we found from document results would include an outlook into the future rankings of happiness for countries within Africa. Andrew E. Clark suggested in World Happiness Report 2016, that the possibility of countries in Africa rising in the ranks of happiness could be fueled by social aspects rather than economical. This is one point that although has merit, we could not support fully. We concluded that social factors are valuable in determining differences between country rankings when they are similar in economic standing however, economic factors played the largest role in determining a higher rank of happiness.

**Deployment**

There are various ways our dataset can be utilized and deployed. In our deployment analysis, we decided to break it down by who will use this information, and how they will use this information. There are issues and important ethical considerations when considering our results. Also, we will discuss the potential risks and backlash that will occur due to our results. First, we will discuss who this data is most vital.

Who will find the results of our analysis the most valuable? Well, we thought about this hard and narrowed it down to the most important people. This includes elected government officials and candidates running for office. Also, the World Health Organization and the World Trade Organization. Global citizens will find our results very informative and knowledgeable. Now that we have identified who will find our results most valuable, we can discuss how they can use this information.

How will these identified people use the results of our analysis to aid them? First, candidates running for office can use these results in a given country to point out what they will focus on to drive their country to the top. For example, our analysis has identified why happiness score is low or high in each country. also, why this country has been given this specific happiness score. Given this, elected officials can discuss their policy to improve the factors that the country is lacking in from our analysis. To be more specific, the U.S. is ranked 13th in happiness. When Donald Trump was running for office, one of his main goals is to strengthen the U.S. economy. He can mention the 2016 dataset results, how we are ranked 13th, what he will do to move the U.S to the number one spot and tie it into his “Make America Great Again” slogan. If we increase our economic power with his policy, America will be happier and move from its 13th overall spot in the world. This would make for a stronger economy, happier citizens, and better poll ratings. This is one example on how candidates running for office can use our analysis.

Government officials already elected in office can use this analysis as well. More specifically, the can use our descriptive analytics to see how their Health, GDP, and Government Corruption is doing. We have seen a positive correlation with an R squared of .7 between Health and GDP. Therefore, if a government official notices their health care is insufficient given our descriptive analytics, they can provide a policy to invest more in health care for their nation. They can use our analysis to prove that investing in health care is a smart option because their GDP (per capita) will also increase as well as citizen’s overall happiness due to a positive correlation from our descriptive analysis. Therefore, government officials can know what to focus more of their attention to in their country given our happiness analysis. This will help government officials to be able to provide a better life for their citizens. Not only will government officials find this information valuable but sectors of the United Nations will as well.

The World Health Organization and World Trade Organization will also find the results of our analysis interesting. The World Health Organization will be able to see the effects of health in various countries and what other variables as well as factors of life the effect on people. Also, the World Trade Organization can see what other factors are leading to declining GDP’s our prospring GDP’s. Companies can also use the results of our dataset when determining where to locate a new business globally. Globalization is becoming a big trend in business. This dataset will provide extensive knowledge in where location a business would be smart and where to not locate a business. Also, for businesses already in the global market, they can discover why they’re business isn't doing so well in a certain country. Now that we have covered who and how the use of our analysis will be used, we must cover risks.

There are potential issues, risks, and backlash we need to consider while deploying our dataset analysis. One thing we have to consider is that people might not believe in our dataset and the results we have obtained. They may be skeptical; however, we must be ethical when we release this information to the public. For example, some countries might ban their citizens from obtaining this type of information. We need to research before we release our results publicly on a global scale for all countries citizens to obtain.

There is also potential backlash between the major people we identified in our deployment discussion from above. For example, if a citizen sees that their happiness is correlating with government corruption, they may protest and start a revolution. These results show that what the government is doing affects citizens lives. Therefore, they may demand a new government in pursuit of better happiness, a stronger GDP, better health care, and less corruption. This could cause several negative circumstances, the worst possibility one could think of is war. We will have to take all of this into consideration when publicly deploying our analysis.