BUDT733: Data Mining for Business

What makes us Happy

By
Team 7

Dan Curtis
Lynn Foo
Prab Goriparthi
Bakta Salla
Rob Whitener

Fall 2009

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

Name	Signature
Lynn Foo	
Dan Curtis	
Rob Whitener	
Prab Goriparthi	
Bakta Salla	

Executive Summary

A simple question by the Gallup world poll posed to people from 138 countries, “Here is a picture of a ladder, suppose that the bottom represents the best possible life and the top the worst possible life. Where on this ladder would you place your current life??

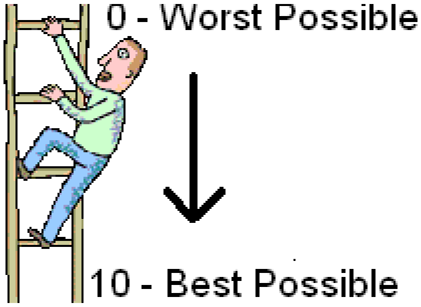


Figure 1. cartoon of scale used by Gallup with ladder going down

The goal of this study is to understand the factors that lead to the inequality of happiness by explaining the factors that drive the happiness scores in the Gallup poll. The results can be used by governments and corporations to guide economic and political policies. Currently, happiness is equated with having money. This is demonstrated in Figure 2 of the appendix which is a map of the world with countries colored by happiness scores. The developed nations (Canada, US, and the EU nations) appear the happiest. The impact of GDP or industrialization on happiness is evident, but it is not the whole story. When the map colored by GDP, Figure 3, is compared to figure 1 there are several countries that are happy despite not being the wealthy as measured by GDP.

The happiness data was cleaned to reduce the number of incomplete and redundant attributes. For example, a simple measure of overall happiness was provided as 42 separate studies. The data covers earlier decades (from the 1940’s) with small sample sizes (<24 countries). The dataset had over a thousand attributes covering almost 200 countries. A contentment measurement of happiness, HappinessBW11Gallup_2006.09, was chosen based on its large sample size (138 countries) and its timeliness (2006-2009). The attributes and countries in our final data set were chosen based on correlation with happiness and the completeness of the data (sample size and completeness of data.) The final data set analyzed have 127 countries and 22 attributes including the y-variable for happiness and 3 PCA attributes.

We chose a logistic regression model that showed happiness is driven mainly by

1. Agrarian Share of the GDP for a country
2. Environmental Performance Index (EPI)

All of our models agreed that countries with higher percentage share of Agrarian GDP aren’t as happy as industrialized nations. In addition, countries with strong environmental performance scored higher in the measurement of happiness.

It is evident from our analysis that a balance between Industry and Environment is the key to happiness. Our analysis kept showing that political stability is a necessary factor in achieving high happiness scores. We recommend that countries and organizations place emphasis on environmental factors to balance advances in industrialization to increase the happiness of their populations.

Technical Summary

Data Source: Our primary data sources were the following:

The Original SPSS file sent by Veenhoven, R., *World Database of Happiness*, Erasmus University Rotterdam. Additional information related to the data file is available at: http://worlddatabaseofhappiness.eur.nl

The World Factbook 2009. Washington, DC: Central Intelligence Agency, 2009. https://www.cia.gov/library/publications/the-world-factbook/index.html
United Nations Human Development Report (http://hdr.undp.org/en/statistics/data/)

We merged the data from all these sources by countries and eliminated those countries that didn't have the Happiness Index information.

After the cleanup we were left with 127 countries (Rows) and 22 columns (1 Y variable, 3 PCA components that captured the political attributes and 18 variables representing Economic and Demographic Indicators. The list of these columns/attributes are shown in Exhibit B

Missing Values: For the missing values, we computed the average by region and filled the missing values with the regional averages. For a few attributes we felt that averages were not the right measure so we filled logically (e.g. Hindu Religion in a predominantly Muslim or Catholic nation was filled with the value of 1)

Outliers: We didn't find any outliers as most of the values were within a given range

Data Exploration: We plotted several graphs in Spot fire (a few graphs shown in Exhibit C). Some attributes displayed a distinctively positive/negative correlation and most others were in a cloud form.

Data Reduction: We grouped the data into categories such as Political, Economic and Demographics. We ran PCA on these groups separately and analyzed the output. In case of Economic and Demographics we could see that a few a attributes/variables contributed significantly to the overall data (the proportion or the weights were significantly larger or smaller than the others) so we kept those attributes and eliminated those attributes that had a very low weights (between -0.0002 and 0.0002). In case of Political attributes, we were unable to observe such distinct separation, so we took the three PCA components that explained 84% of the variation. We also decided not to partition the data as we were left with very few rows for analysis.

Models: We ran the following models on the cleaned up data to cross check the performance of each model:

Classification Tree
Logistic Regression
Discriminant Analysis and
Linear Regression

We ran Classification Tree and Linear Regression even though they were not suitable (small data size for classification tree and Categorical Y Variable for Linear Regression) in order to validate our end model. The performance of the above four models are as shown below.

	Classification Tree	Linear Regression	Logistic Regression	Discriminant Analysis
Number of Variable s	AgrarianShareGDP_05 EPI_06 TelephoneLines_2000 GovIntervention2_2006 FreeEconIndex2_2007	AgrarianShareGDP_05 EPI_06 RuleofLaow_BS2008_Rank PCA1 PCA2	AgrarianShareGDP_05 EPI_06	All Variables Listed in Exhibit B
Multiple R-Sq		0.753883863	0.5001415	
PctError	12.60	0.494143927 (RMSE)	12.60	13.39
Cut off Prob.	0.5		0.5	

Based on the performance measures and the end goal, we selected Logistic Regression to be the best fit for explaining the happiness in people.

Logistic Regression Model to explain the happiness: In the end the following attributes were used to run the logistic regression

EPI_06 (Environmental Performance Index, 50% of this index represents Environmental Health factors such as: environmental burden of disease, water health: sanitation, drinking water and air pollution. The remaining 50% of this measure represents *Eco-system vitality* factors such as: air, water, biodiversity , use of natural resources and addition to climate change). Value range is from 0 to 100, value range in the data set is between 25.7 and 88. Higher values represent a better environment.

As of 2006, A Higher index measure represents a higher happiness. A better, clean and safe environment makes people much more happier than a polluted and unsafe environment.

AgrarianShareGDP_05 (Share of Agriculture – Includes Forestry, fishing, hunting, cultivation of crops and livestock - as a percentage of GDP – Sourced from World Development Indicators 2007, table 6.14. Value ranges are from 0 to 48%)

As of 2005, A country that has a Higher percentage of GDP from agriculture is on average, considered less happier than a country that has a lower percentage of GDP from Agriculture. In other words an Industrialized country on average is happier than a less industrialized country.

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-1.73085833	2.2630105	0.44436225	*
AgrarianShareGDP_05	-0.16430622	0.04305735	0.00013564	0.84848219
EPI_06	0.06760561	0.02974937	0.02305598	1.06994331

Residual df	124
Residual Dev.	84.18389893
% Success in training data	62.20472441
# Iterations used	9
Multiple R-squared	0.5001415

Classification Confusion Matrix		
	Predicted Class	
Actual Class	UNHAPPY	HAPPY
HAPPY	73	6
UNHAPPY	10	38

Error Report			
Class	# Cases	# Errors	% Error
HAPPY	79	6	7.59
UNHAPPY	48	10	20.83
Overall	127	16	12.60

Exhibit A

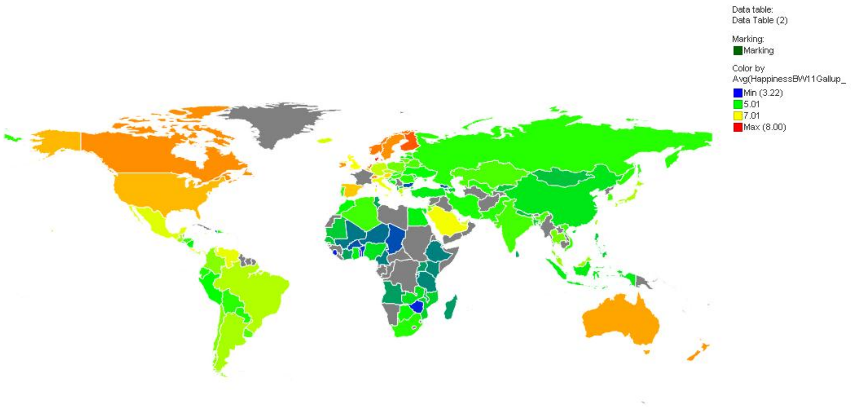


Figure 2: World map showing happiness scores by country. Note: grey countries show missing data in graph)

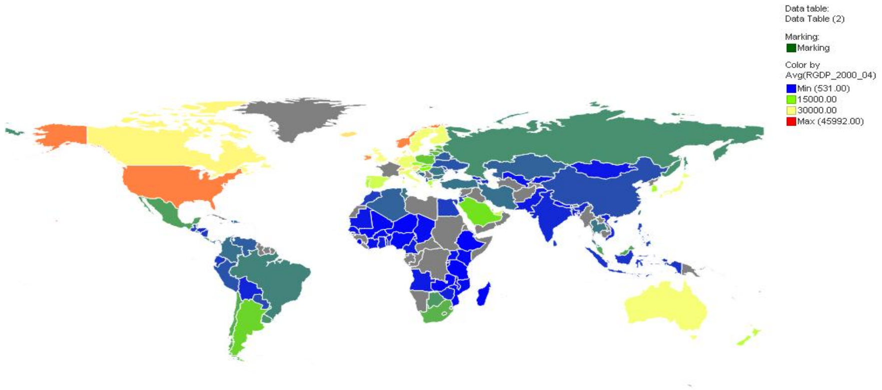


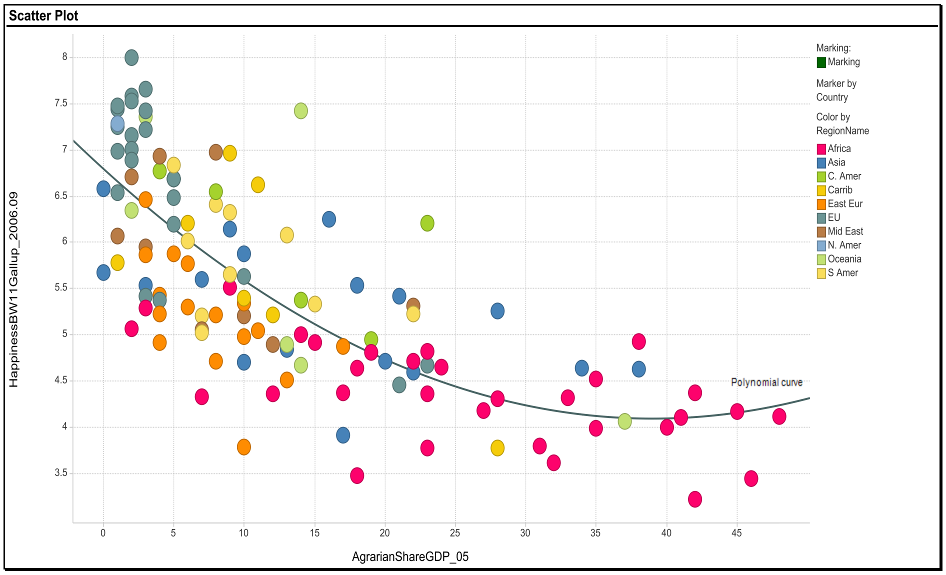
Figure 3. World Map of Real GDP measured in 2004.

Exhibit B

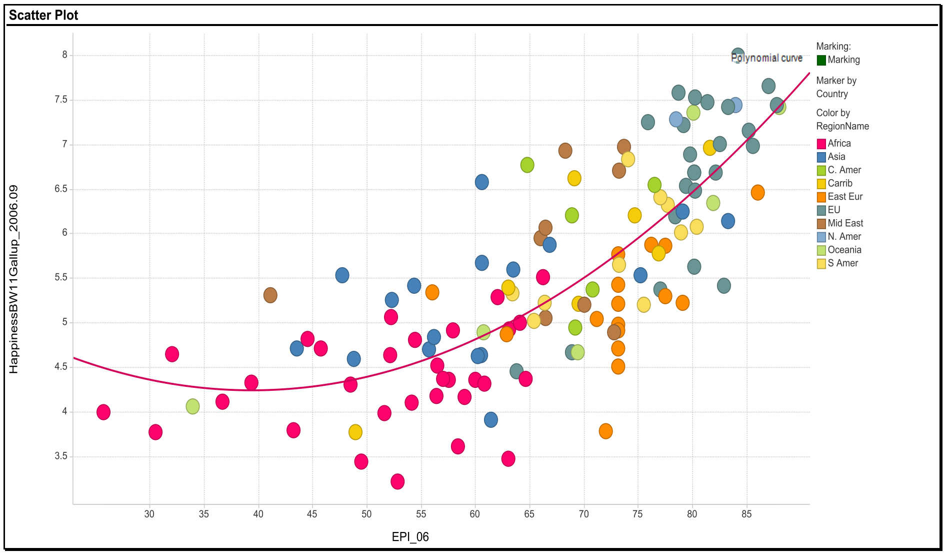
List of Attributes after Clean up

MobilePhones_2005	FreeEconIndex2_2007	Reg_Quality_BS2008%Rank
InternetUse_2005	FreeEcon2Real_2006	Voice_Account_BS_2008
TelephoneLines_2000	EduEnrolGross_2000_04s	RuleofLaow_BS2008_Rank
AgrarianShareGDP_05	EPI_06	GovEffect_BS_2008
FreePress3_00s	BusinessFreedom2_2006	PCA1
GovIntervention2_2006	InvestmentFreedom2_2006	PCA2
FreeTrade2_2006	Poli_Stab_BS_2008	PCA3

Exhibits C



Exploration: Agrarian Share of GDP vs Happiness



Exploration of Environmental Performance Index (EPI) vs Happiness

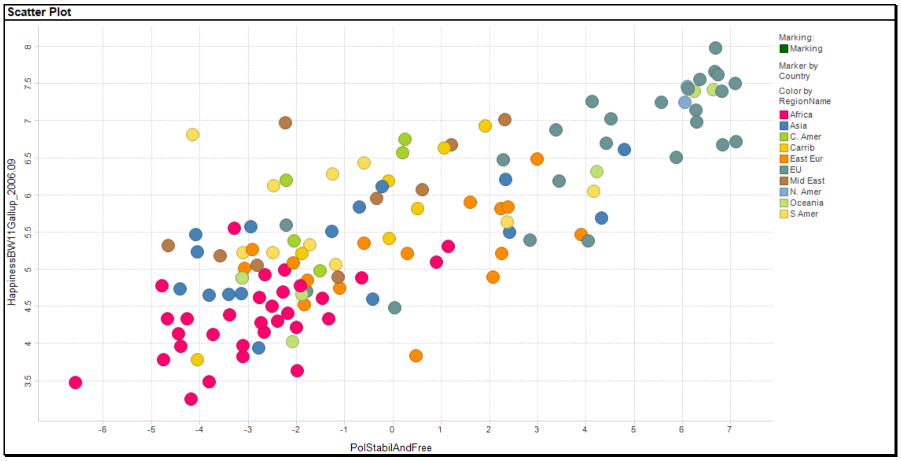


Figure 4: PCA1 (Political Stability and Freedom) v. Happiness