Are we happy with our life?

Xiaoqian Dang

April 2, 2017

Abstract: In this report, I analyze a problem that is related to our emotion: happiness and life. We study the happiness score variation for different countries. We consider many possible factors that might affect our happiness, such as GDP, total income, and even alcohol consumption. Our study was not aimed to a biological interpretation of happy or sad, instead, we studied what is the external or political reason that makes people have different emotion around the world. In order to find out the importance of factors, we use a linear regression analysis and inferential statistical methods. At last, we compared the effect of different algorithms on the predictions.

1. Introduction

Are we happy? This is a question we ask all the time. Our society has been trying to answer this question for thousands of years. This question is so interesting that people developed many different ways to answer it: Philosophical, biological, and until recently economic and sociological. In this report, we are not trying to answer this question by philosophical or biological point of view. Instead, we focus on sociological and economic factors. What makes us happy? What is the reason? Income? Peace? Alcohol? Or even smoking? Does our cultural background also affects our emotion? In this project, I would like to do some research on what are the features that make us happy and which is the most important. Additionally, I would like to investigate some interesting correlations among different factors. We can apply statistical method to answer those questions. In addition to these possible questions, we can also try to answer questions such as: can we build a linear regression model to predict if people are happy or not. Of course, this is not the only way to explore this question. Some other questions that could also be investigated are: Do different countries have different happiness levels? Are people at different periods of time have the same level of happiness? There are almost infinite possible questions to ask, however, I will focus only on a dataset that includes macroeconomic parameters so that the model calculation is possible.

My report is related to a 'big' topic so that there are many different algorithms to perform the evaluation of this project. But for my interest of study, I would like to perform a regression study; my focus is on the relation between happiness and total income, country, time, and other social problems. What I want to determine is what are the important factors that determine our happiness. Can we make any predictions based on the model? Or can we classify the countries in the world based on their happiness index (of course, we need to find out the mathematical definition of this index. This is not too difficult to do since there have been many researches on this topic)? The difficulties I might face are: the data that are going be used to answer this question are distributed in different data set. So in order to have all the date we need, I have to spend lots of time on data cleaning. Also, the data might be noisy, that will make the analysis data difficult. For the purpose of study, some statistical analysis is also necessary.

As I mentioned in the previous section, for this problem, we can try to answer the questions in four aspects: What is the culture and geographic distribution of the happiness? How does this value vary with time? How is the happiness distributed for different continents? What is the core factor that significantly affects our happiness? What is their relation? Can we make a prediction of happiness based on it? In order to answer this question, it is impossible to go over all possible factors that might affect the result, so I want to limit our study on some factors, such as: (1). The total income of the country (here, people often use two index to show the total income: GDP per capita or log GDP per capita. Log of Per capita GDP = Log (Total GDP/ Population)); (2). Food consumption. Here we only restrict ourself on wheat and vegetables. (3). The working hours for different country; can we categorize the happiness for different countries? We can analyze the correlation between happiness and some other social problems: alcohol/tobacco consumption, suicide rate, etc. Here, we can compare the alcohol consumption data with the happiness index. In the next section, I will list all the features I am going to use throughout our investigation.

The structure of this report is as follows: in the second section, I discussed the original dataset and how I cleaned the dataset in order to perform our model calculation. Along with the description of data wrangling, I also give some qualitative study of the dataset, such as distribution of happiness score; correlation between Happiness score and other features. In the third section, we first consider a full feature model. Based on that model, I performed feature selection in order to reduce the number of features for further study. The goal of this step is to find out a smaller dataset to make an appropriate prediction and minimize the possibility of overfitting the data. In the last section, I finalize the model and compare the prediction between different machine linear algorithms.

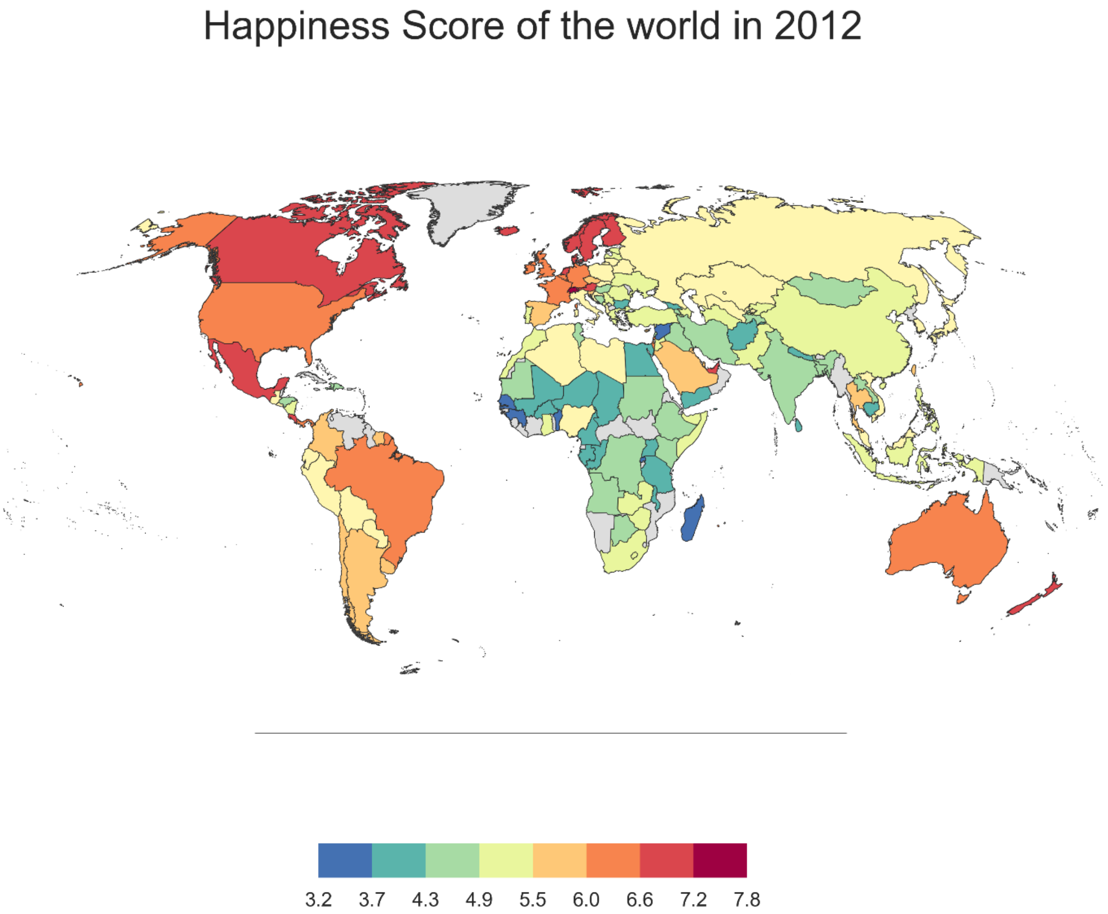


Figure 1 The world happiness score of year 2012. Different colors show different values of happiness score. The redder the color the happier the people in that country. This map shows the strong relation between the economic features with the happiness score.

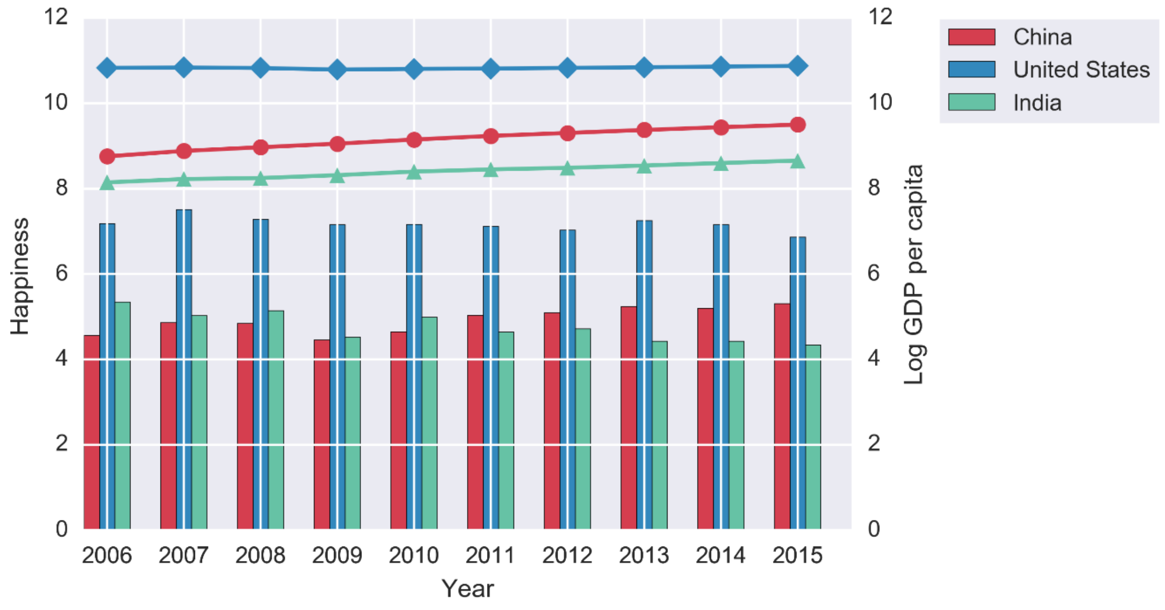


Figure 2 Distribution of happiness score and logarithm GDP at different years. The straight line in the upper part of the picture is the GDP distribution. The barplot is the happiness score.

2. Data preparation and data cleaning

In this section, I will discuss the process of preparing the dataset I use for our machine learning study. This section is divided into two different parts. In the first part, I will point out the data resources and explain why I choose this dataset and what the data looks like. Also, a very intuitive description of the data is shown right after it. In the second part of this section, I will show and discuss the correlation between each feature.

2.1 Getting to know the dataset

As I have mentioned in the introduction section, most of our data is from the Gallup world happiness poll[1], which they collected the data about happiness, GDP by making survey. Their

[1] http://worldhappiness.report/

original dataset starts from 2006 to 2015. But for our purpose of study, we only focus on the year 2012, which has the most complete data values. As the first glance, I put some basic information about happiness in Figure 1, Figure *2*. These two figures show the geographic relation and dependence of happiness score on GDP clearly. Some basic information we can get from these plots are: the happiness distribution is not evenly distributed around the world; GDP is highly related to the happiness score, but not the only reason.

In order to make our dataset more realistic, I collected data from other sources. The final dataset could be found in my GitHub repository[2]. It has 139 countries and regions around the world, and 19 features come along with the happiness score. Due the limit of the space, I will not post the whole data frame in this report. Instead, I will list the name and meaning of each feature in the following list:

1. **Log GDP per capita**: This is the quantity that properly describes the GDP of a country. The definition is: log(GDP_per_capita)=log(GDP/population)

2. **Confidence in national government**: The higher this value, means higher reliable of the government. The data range is from 0.0 to 1.0.

3. **Social support**: is the national average of the binary responses (either 0 or 1) to the 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?"

4. **Healthy life expectancy at birth**: The time series of healthy life expectancy at birth are calculated by the authors based on data from the World Health Organization (WHO), the World Development Indicators (WDI), and statistics published in journal articles.

5. **Generosity**: This is the national average of response to the GWP question "Have you donated money to a charity in the past month?" The data is calculated on GDP per capita. It shows the percentage of money is donated to charities. The negative value means the opposite meaning.

[2] https://github.com/dangspin/SpringBoard_Projects/tree/master/final_project

6. **GINI index (World Bank estimate), average 2000-13**: from the World Development Indicators (Last Updated: 22-Dec-2015). The variable labeled at the source as "GINI index (World Bank estimate)", series code "SI.POV.GINI". The average does not imply that a country has the Gini index in all years in that period. In fact, most do not.

7. **alcohol**: Alcohol, recorded per capita consumption (in liters of pure alcohol)

8. **food**[3]: Discussing the absolute food consumption is not quite meaningful since it is confusing to say someone has more food than others. Instead, here we use the Kcal per capita to show the total food consumption, which indicate how much calorie a person need for one day.

9. **Expenditure_on_health**[4]: Total expenditure on health per capita in U.S. dollar. Total health expenditure is the sum of public and private health expenditures as a ratio of total population.

10. **homicide**[5]: Rate per 100,000 population. "Intentional homicide" is defined as unlawful death purposefully inflicted on a person by another person

11. **child_mortality_rate**: Deaths under age of five per 1,000 live births. Both sexes combined. Probability of dying between birth and exact age 5. It is expressed as deaths per 1,000 births.

12. **income**: Gross National Income- GNI per capita (in U.S. dollar)

13. **university_enrollment_rate**[6]: Percentage of population enrolled in university.

14. **Expenditure_on_education**[7]: Public spending on education (% of GDP) - countries ranking (% of GDP)

[3] https://knoema.com/atlas/topics/Agriculture/Food-Supply-Total-Energy-kcalcapitaday/Total-food-supply

[4] https://knoema.com/atlas/topics/Health/Health-Expenditure/Health-expenditure-percent-of-GDP

[5] https://knoema.com/atlas/topics/Crime-Statistics/Homicides/Homicide-rate

[6] https://knoema.com/atlas/topics/Education/Tertiary-Education/Gross-enrolment-ratio

[7] https://knoema.com/atlas/topics/Education/Expenditures-on-Education/Public-spending-on-education-percent-of-GDP

15. **visitor_per_hectare**: This is the average of total visitor numbers per year. The average is= Total visitors/ land area. The unit is hectare= 10000 m^2

16. **unemployment_rate**: (%) of unemployed labor force.

17. **economic_freedom_index**[8]: Economic Freedom. (score 100 represents the maximum freedom). Economic freedom is the fundamental right of every human to control his or her own labor and property[9]. In an economically free society, individuals are free to work, produce, consume, and invest in any way they please, with that freedom both protected by the state and unconstrained by the state. In economically free societies, governments allow labor, capital, and goods to move freely and refrain from coercion or constraint of liberty beyond the extent necessary to protect and maintain liberty itself.

18. **total_visitors**[10]: Total number of visitors this year

19. **Land_area**[11]: The land area of this country in the units of km^2

20. **suicide**: The suicide rate per 100,000 population

Now let's take a look at some of the features' properties. The first is the dependent variable of our model, the happiness score.

[8] https://knoema.com/atlas/topics/World-Rankings/World-Rankings/Index-of-economic-freedom

[9] https://en.wikipedia.org/wiki/Economic_freedom

[10] https://knoema.com/atlas/topics/Tourism/Key-Tourism-Indicators/Number-of-arrivals

[11] https://knoema.com/atlas/topics/Land-Use/Area/Country-area

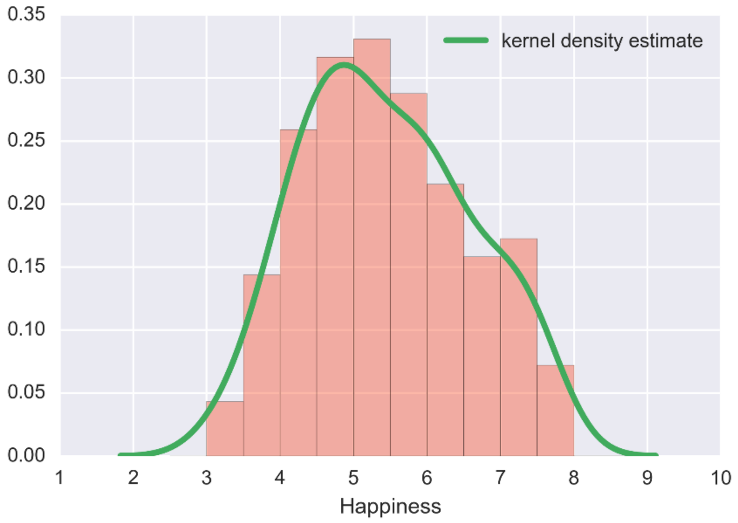


Figure 3 The histogram plot of the distribution of happiness score. The green curve is the calculated kernel density distribution

The distribution of the happiness score around the world is similar to the normal distribution with mean value 5.45 and standard deviation 1.128. This makes sense, because as a social phenomenon, what we can expect is that not every are really happy with their life, but other hand, not everyone feels sad about their life. The normal distribution is an appropriate distribution to describe it.

Knowing the correlation between different features and the happiness score is also important and interesting since this information could provide an impression of the relationship within the dataset. A full investigation is not necessary since scatterplot only gives us a intuitive way to understand the data, so instead, I only consider two interesting features in our dataset: 1. LogGDP and 2. Alcohol consumption.

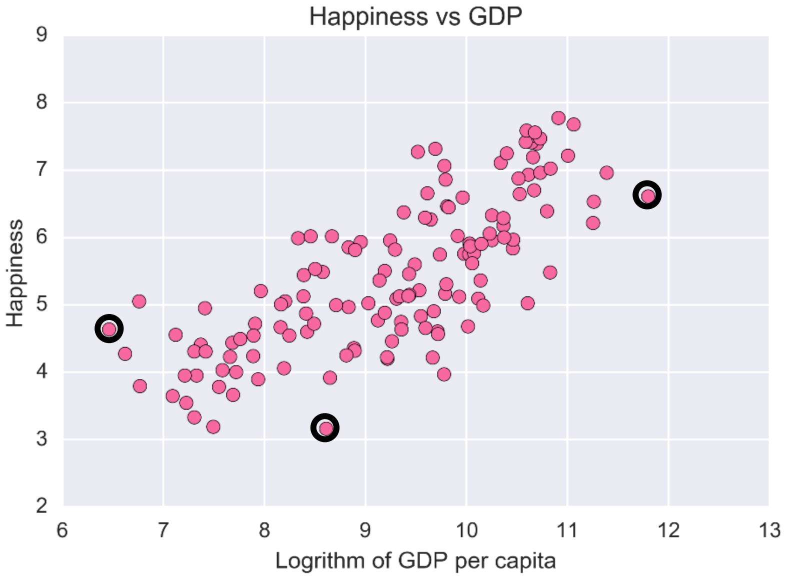


Figure 4. The scatterplot between Logarithm of GDP versus happiness score.

In Figure 4, we plot the correlation between Logarithm function of GDP versus the happiness score. The general tendency of the data shows a strong linear correlation, but with a high variation perpendicular to the linear relation direction. However, if we want to ask how strong GDP will affect the happiness and how GDP compares with other factors, we need to build a regression model. There are some outliers that are quite interesting. These countries are highlighted in Figure 4 as indicated by black circle outlines.

The Syrian Arab Republic. The country has a fair high value of GDP (8.6 as Figure 4 shows) but the happiness score is lower than the average of that level of GDP. This country has the lowest happiness score, which indicates that there are some other factors affect their happiness. Qatar, on the contrary has highest value of GDP (11.8) but the people are not the happiest country. It also tells us that a useful model need to include other factors. At last, the democratic republic of Congo, also as known as Congo-Kinshasa. The people in this country has the lowest GDP per capita, but they are really happy with their life!

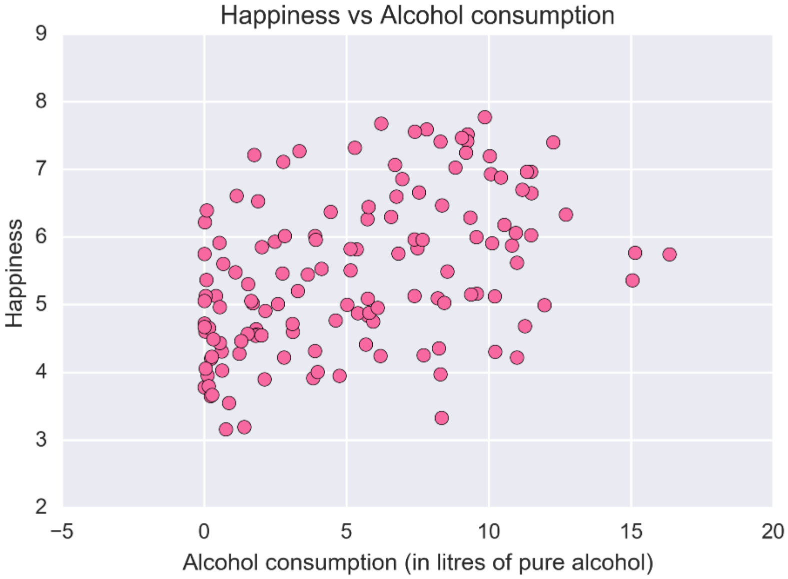
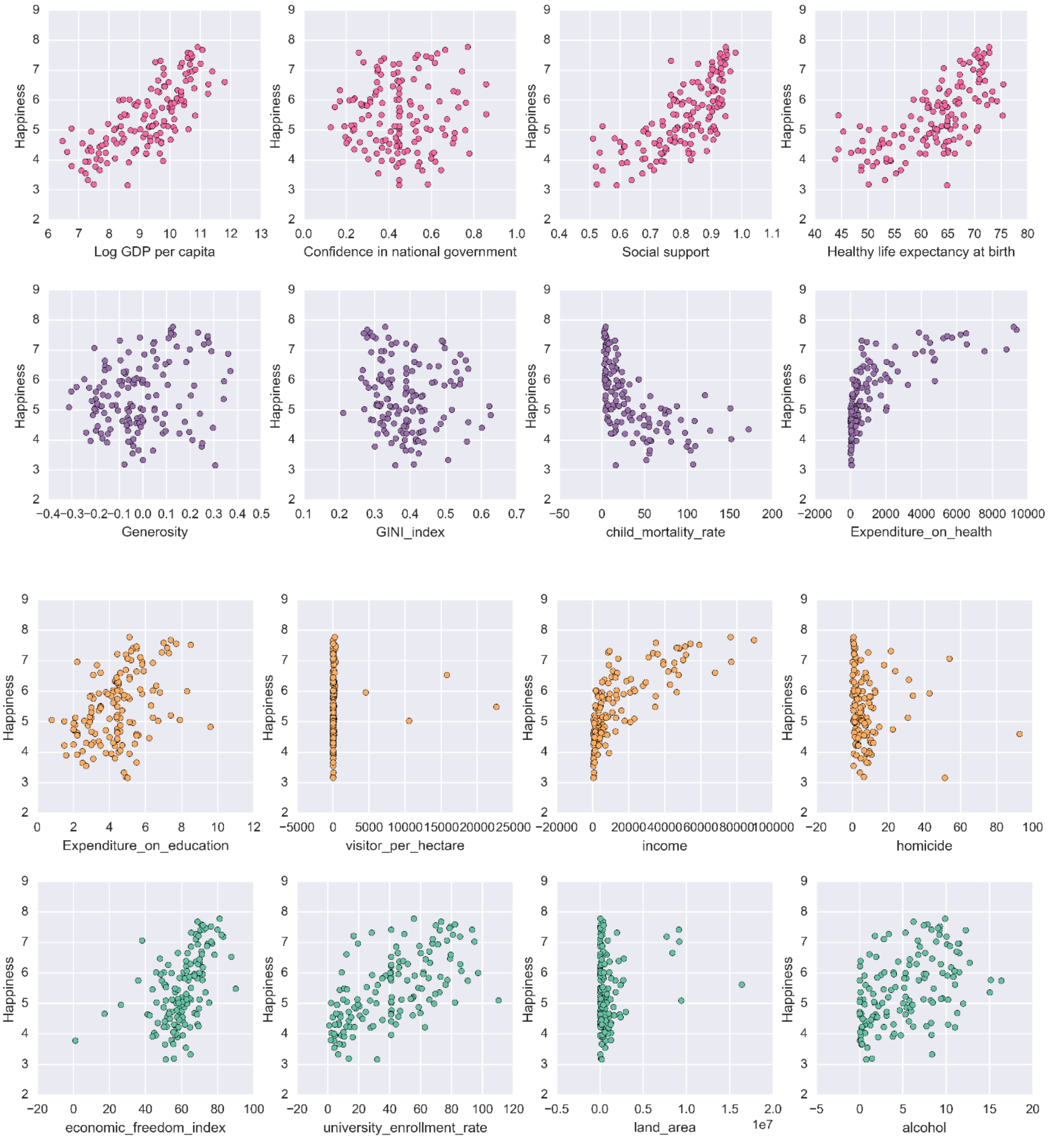


Figure 5. The scatterplot between alcohol consumption and happiness score.

This scatterplot shows no relation between alcohol consumption and Happiness score. Our conclusion is: in some countries, people use alcohol to make them happier, but in some other country, alcohol is not that important. But in some country, the more people use alcohol, the unhappier they are. In section 3, we will see that this feature is not really important in our model, and we can remove it without hurting the model prediction result.

2.2 Features correlation

After investigating our dataset, the next step is to figure the correlation among different features. The best way to understand the correlation relation is by looking at the scatterplot matrix that shows the relation between features and happiness score.



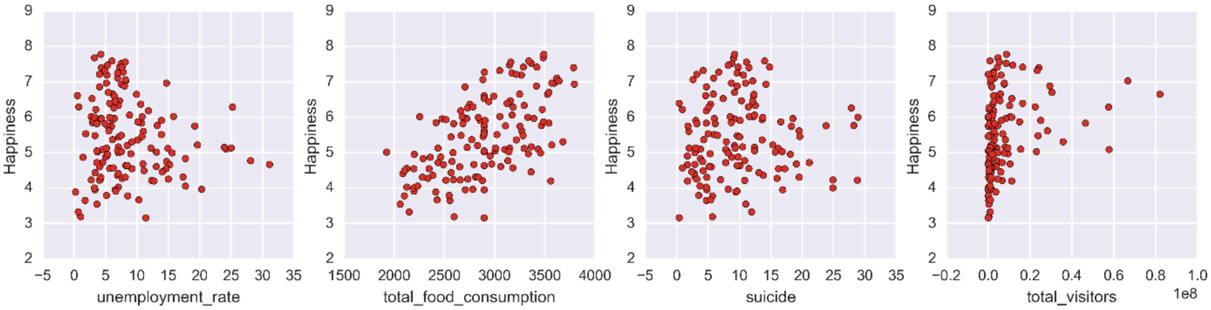


Figure 6 The scatterplots between different features and happiness score. We have total 20 features in our dataset.

	Happiness
Happiness	1.000000
Log GDP per capita	0.753528
Confidence in national government	0.026958
Social support	0.723328
Healthy life expectancy at birth	0.699890
Generosity	0.215839
GINI_index	-0.124207
Expenditure_on_education	0.419676
homicide	-0.094060
economic_freedom_index	0.502716
university_enrollment_rate	0.617572
alcohol	0.443823
unemployment_rate	-0.126270
total_food_consumption	0.592544
suicide	0.036321
total_visitors	0.295438
log_child_mortality_rate	-0.696524
log_Expenditure_on_health	0.810017
log_income	0.806295
log_visitor_per_hectare	0.353837

Table 1 Correlation between happiness score with other features.

Figure 6 Shows that GDP, total income, social support, healthExp, etc. are highly correlated to the happiness score of the country. However, the consumption of alcohol, suicide, and homicide rate are not strongly predictive of happiness. This is quite counterintuitive since what we might expect

is that the more we consume alcohol, the happier we might be. This first step investigation strongly indicates us, a mathematical model is important for our model calculation and prediction.

In Figure 6, I consider the correlation relation only for the original dataset, however, the features child, expenditure_on_health, income and visitor_per_hectare strongly indicate the exponential relationship, so that in Table 1, we considered the logarithm of these features in order to increase their correlation.

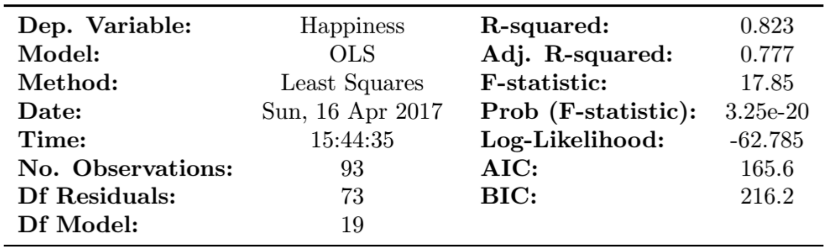
3. Feature selection and model training

After we finished the previous section, we have a relatively clean dataset without any null data values. Now the dataset is ready proceed to the training process. In this section, I will perform the feature selection process by using Python statistical and machine learning package Statsmodels, which provides us a very detailed information about model criteria and training parameters. Our investigation is divided into three different steps. First, I will consider the model training for the full-feature dataset. Then a greedy feature selection algorithm is performed in order to pick up the most important features in order to prevent overfitting. As the last step, we tune the model parameters to figure out the most reasonable feature selection.

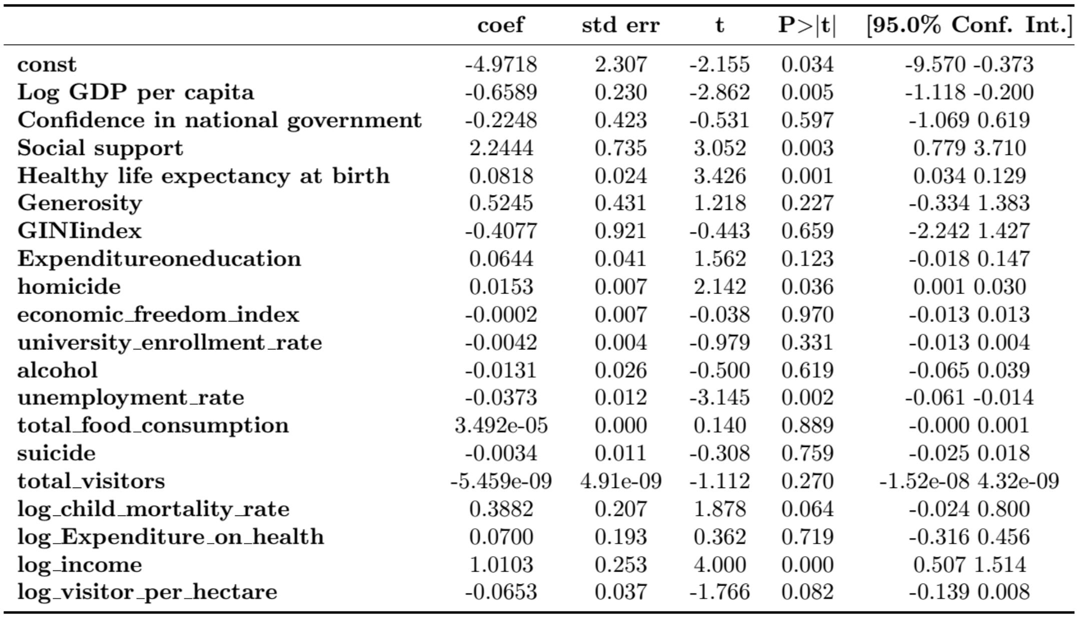
Table *2* shows the result of our multivariable linear regression model, for which it calculates the coefficients of the model by minimizing the sum of squared residuals (SSR):

$$SSR = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{1}$$

Our results show a relative high R^2 and Adjusted R^2 values, which tell us that our model fits the training dataset pretty well. However, due to the fact that many of the features are strongly correlated to each other, it is not wise to choose this model as our final machine learning algorithm for the problem, so that this result indicates that a feature selection process is necessary. At the same time, the feature selection might help us prevent overfitting, which will destroy the prediction accuracy.



Dep. Variable:	Happiness	**R-squared:**	0.823
Model:	OLS	**Adj. R-squared:**	0.777
Method:	Least Squares	**F-statistic:**	17.85
Date:	Sun, 16 Apr 2017	**Prob (F-statistic):**	3.25e-20
Time:	15:44:35	**Log-Likelihood:**	-62.785
No. Observations:	93	**AIC:**	165.6
Df Residuals:	73	**BIC:**	216.2
Df Model:	19		

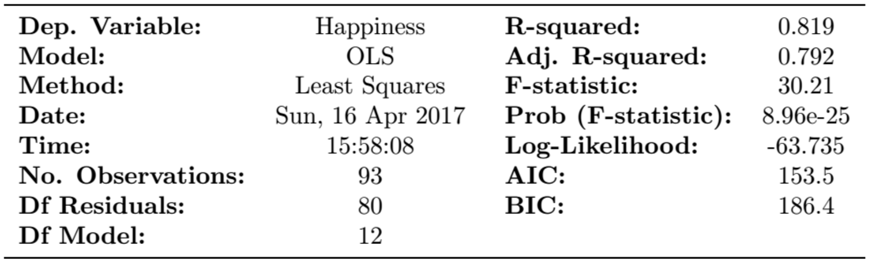


	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	-4.9718	2.307	-2.155	0.034	-9.570 -0.373
Log GDP per capita	-0.6589	0.230	-2.862	0.005	-1.118 -0.200
Confidence in national government	-0.2248	0.423	-0.531	0.597	-1.069 0.619
Social support	2.2444	0.735	3.052	0.003	0.779 3.710
Healthy life expectancy at birth	0.0818	0.024	3.426	0.001	0.034 0.129
Generosity	0.5245	0.431	1.218	0.227	-0.334 1.383
GINIindex	-0.4077	0.921	-0.443	0.659	-2.242 1.427
Expenditureoneducation	0.0644	0.041	1.562	0.123	-0.018 0.147
homicide	0.0153	0.007	2.142	0.036	0.001 0.030
economic_freedom_index	-0.0002	0.007	-0.038	0.970	-0.013 0.013
university_enrollment_rate	-0.0042	0.004	-0.979	0.331	-0.013 0.004
alcohol	-0.0131	0.026	-0.500	0.619	-0.065 0.039
unemployment_rate	-0.0373	0.012	-3.145	0.002	-0.061 -0.014
total_food_consumption	3.492e-05	0.000	0.140	0.889	-0.000 0.001
suicide	-0.0034	0.011	-0.308	0.759	-0.025 0.018
total_visitors	-5.459e-09	4.91e-09	-1.112	0.270	-1.52e-08 4.32e-09
log_child_mortality_rate	0.3882	0.207	1.878	0.064	-0.024 0.800
log_Expenditure_on_health	0.0700	0.193	0.362	0.719	-0.316 0.456
log_income	1.0103	0.253	4.000	0.000	0.507 1.514
log_visitor_per_hectare	-0.0653	0.037	-1.766	0.082	-0.139 0.008

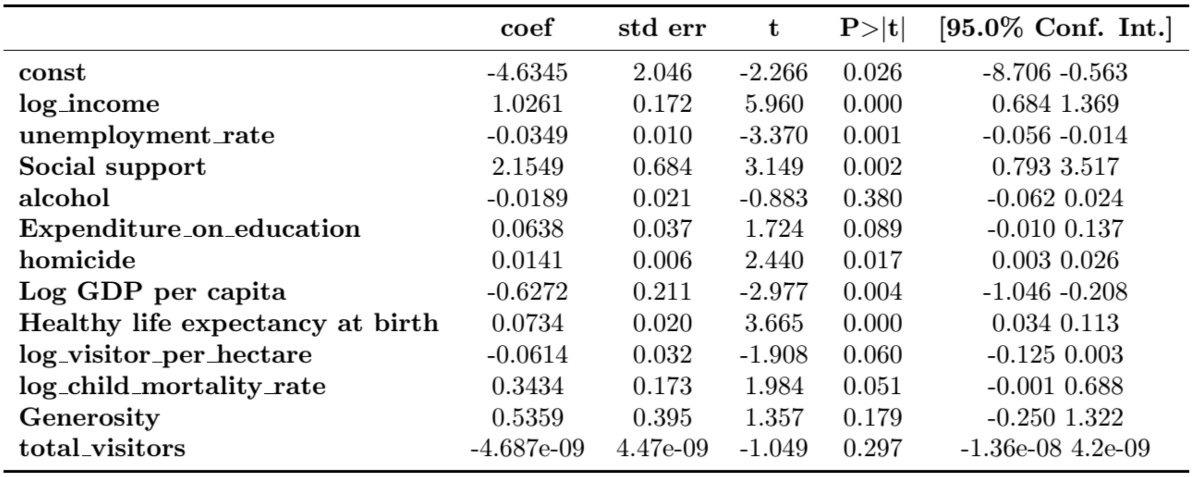
Table 2 The linear regression result of the full-feature model. The first part of the table gives us the basic information about the model. The rest of the table are the value and statistical information for each coefficient.

There are many possible ways to perform feature selection, such as PCA, decision tree algorithm, or even a full combination of feature selection. All of these selection algorithms have their own advantages and disadvantages, i.e. the PCA method is useful for dimension reduction and feature selection, but it cannot give us a direct meaning of the new principle axis. Because of this, in our study, I perform forward feature selection algorithm, which is basically a greedy algorithm. This method is relatively fast but we have to pay for the price that it might not give us the best coefficient estimates.

The final result is that only 12 features are selected as model independent variables. The calculated result of our new model is showed in:



Dep. Variable:	Happiness	**R-squared:**	0.819
Model:	OLS	**Adj. R-squared:**	0.792
Method:	Least Squares	**F-statistic:**	30.21
Date:	Sun, 16 Apr 2017	**Prob (F-statistic):**	8.96e-25
Time:	15:58:08	**Log-Likelihood:**	-63.735
No. Observations:	93	**AIC:**	153.5
Df Residuals:	80	**BIC:**	186.4
Df Model:	12		



	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	-4.6345	2.046	-2.266	0.026	-8.706 -0.563
log_income	1.0261	0.172	5.960	0.000	0.684 1.369
unemployment_rate	-0.0349	0.010	-3.370	0.001	-0.056 -0.014
Social support	2.1549	0.684	3.149	0.002	0.793 3.517
alcohol	-0.0189	0.021	-0.883	0.380	-0.062 0.024
Expenditure_on_education	0.0638	0.037	1.724	0.089	-0.010 0.137
homicide	0.0141	0.006	2.440	0.017	0.003 0.026
Log GDP per capita	-0.6272	0.211	-2.977	0.004	-1.046 -0.208
Healthy life expectancy at birth	0.0734	0.020	3.665	0.000	0.034 0.113
log_visitor_per_hectare	-0.0614	0.032	-1.908	0.060	-0.125 0.003
log_child_mortality_rate	0.3434	0.173	1.984	0.051	-0.001 0.688
Generosity	0.5359	0.395	1.357	0.179	-0.250 1.322
total_visitors	-4.687e-09	4.47e-09	-1.049	0.297	-1.36e-08 4.2e-09

Table 3 The linear regression result of the 12-feature model. The first part of the table gives us the basic information about the model. The rest of the table are the value and statistical information for each coefficient.

If we compare with table 1 and 2, we will see that our 12-feature model has a lower R^2, higher adjusted R^2. At the same time, our new model has a lower AIC number, which is defined as:

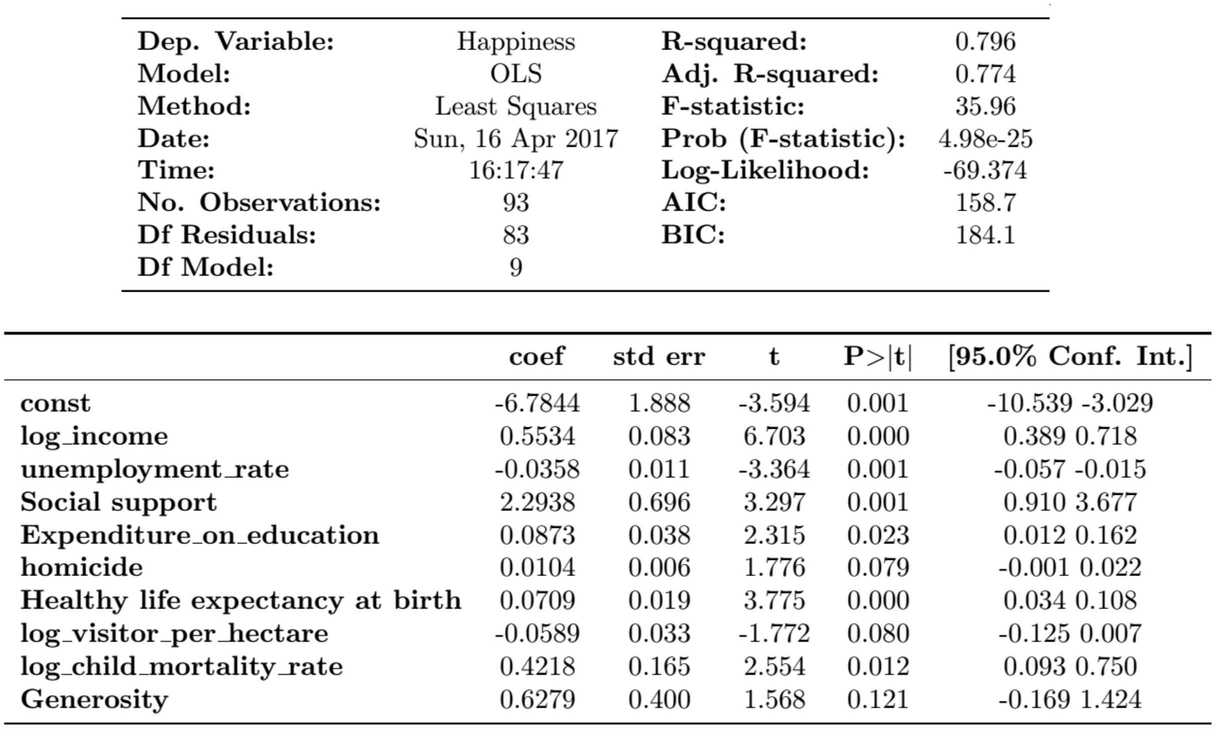
$$AIC = 2k - 2\ln(\hat{L}) \tag{2}$$

Where $\hat{L}$ is the maximum log likelihood function and *k* is the number of features in our model. This is a good sign for our new model, but it has fewer features than the original one and prevents overfitting.

It looks like we have very good results and our model has a promising feature set. But if we look at table 2 carefully, we will see that there are some problems worthy of further investigation. The feature 'GDP' has a negative coefficient in our multivariable linear model. This is counterintuitive result since, in the previous section, the correlation picture and matrix show the opposite result. The possible reason why we have a negative value that should be a positive number is due to the strong correlation between features. This is the phenomenon called multicollinearity, where the strong correlation between features could make the sign of the coefficient change. In order to avoid this dilemma, I would sacrifice the precision of our new model and remove this, so that our model will have a reasonable interpretation.

There is another feature that also need our attention that the feature 'visitor' has a relative high $p-value$, and very small coefficient. The meaning of *p*-value is that if we have a null hypothesis that this feature has no effect on our model. This tells us that we can throw away this feature without affecting our model prediction too much.

As our final result, we keep only 9 features in this model, and the value of each feature is listed in the following table, table 4:



Dep. Variable:	Happiness	**R-squared:**	0.796
Model:	OLS	**Adj. R-squared:**	0.774
Method:	Least Squares	**F-statistic:**	35.96
Date:	Sun, 16 Apr 2017	**Prob (F-statistic):**	4.98e-25
Time:	16:17:47	**Log-Likelihood:**	-69.374
No. Observations:	93	**AIC:**	158.7
Df Residuals:	83	**BIC:**	184.1
Df Model:	9		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	-6.7844	1.888	-3.594	0.001	-10.539 -3.029
log_income	0.5534	0.083	6.703	0.000	0.389 0.718
unemployment_rate	-0.0358	0.011	-3.364	0.001	-0.057 -0.015
Social support	2.2938	0.696	3.297	0.001	0.910 3.677
Expenditure_on_education	0.0873	0.038	2.315	0.023	0.012 0.162
homicide	0.0104	0.006	1.776	0.079	-0.001 0.022
Healthy life expectancy at birth	0.0709	0.019	3.775	0.000	0.034 0.108
log_visitor_per_hectare	-0.0589	0.033	-1.772	0.080	-0.125 0.007
log_child_mortality_rate	0.4218	0.165	2.554	0.012	0.093 0.750
Generosity	0.6279	0.400	1.568	0.121	-0.169 1.424

Table 4. The linear regression result of the *9*-feature model. The first part of the table gives us the basic information about the model. The rest of the table are the value and statistical information for each coefficient. This is our final model.

In order to get to know more about our new model, I plot the residual-fitted plot and actual-fitted plot as reference. The almost evenly distribution of residual-fitted plot indicates that our model fits well. This point could also be verified by the actual-fitted plot. For a perfect linear relation, this plot is along the diagonal direction as the red dashed line shows. In our case, although our data has variation, the general trend is still linear along the 45° line. In the next section, I discuss the machine learning algorithm based on our 9-feature model.

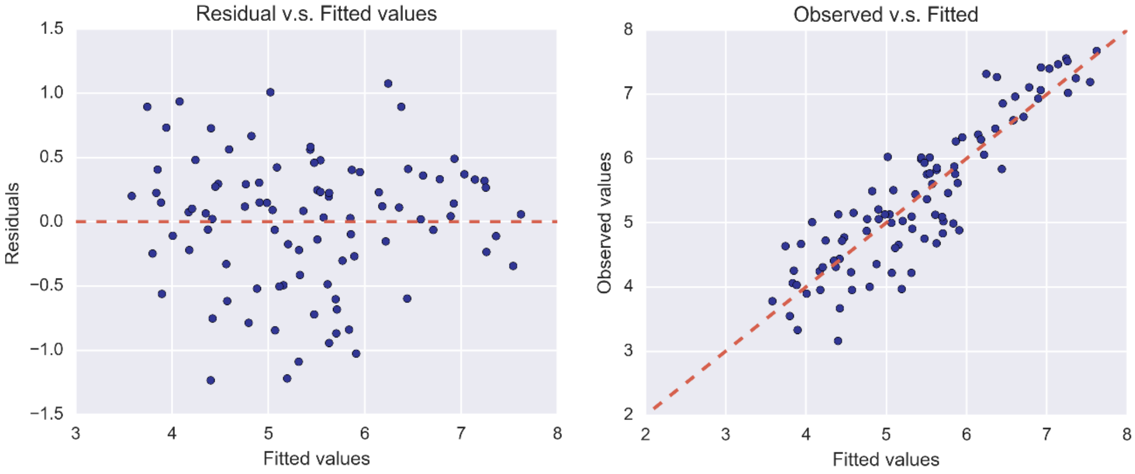
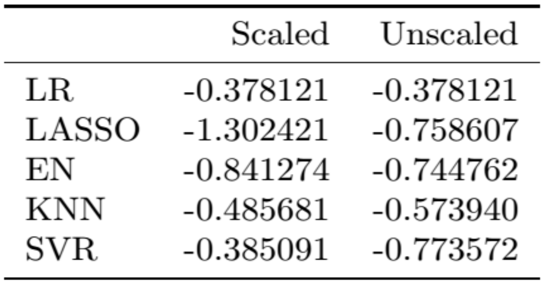


Figure 7 Left: the residual vs fitted values plots. Right: The Observed (Actual) vs Fitted plots. These two plots help us to understand how our model fits the real data.

4. Model selection and parameter tuning

In section 3, we study the feature selection problem, where we learned that the full-feature and 12-feature model although give us a high value of R^2, he multicollinearity prevents them to be interpretable model for future study. As a matter of fact, I manually removed three other features that have higher *p-value* and smaller coefficients. In this section, I apply different machine learning algorithms on the 9-feature model. The goal of this study is to compare the result of different algorithms and convince us that our dataset can give us solid result no matter which algorithm we choose.

The machine learning algorithms for our study are regular linear regression (LR); LASSO; Support vector regression (SVR); K-nearest neighbor (KNN) and elastic net (EN). Since different algorithm has different sensitivity to the dataset, we also perform the calculation for scaled and unscaled dataset. Here scaled dataset means a data distribution that it behaves as Gaussian distribution with $\text{mean} = 0, \text{ variance} = 1$. I list the result in Table *5*:



	Scaled	Unscaled
LR	-0.378121	-0.378121
LASSO	-1.302421	-0.758607
EN	-0.841274	-0.744762
KNN	-0.485681	-0.573940
SVR	-0.385091	-0.773572

Table 5 The comparison between different machine learning algorithms with and without standard scale. Here we use Python scikit-learn package, where the built-in measure of precision is negative mean square error, so that the bigger the value the better the model. Interestingly, the regular linear regression is insensitive to the dataset.

Our model comparison table help us pick up the most appropriate model for our study, they are regular linear regression (LR) and Scaled SVR, which show a similar result. There is one important thing I have to mention that, this scaled SVR calculates by using default parameter. The most accurate SVR requires cross-validation to find out the best soft margin parameter *C* and kernel function.

As the last step of our model study, I apply these two algorithms on the test dataset, which is generated by random train-test split. The mean squared error is used as the model prediction criteria and our result is:

$$\begin{aligned} &LR:\ 0.24 \\ &SVR:\ 0.30 \end{aligned} \tag{3}$$

The difference between these two algorithm is not quite big, so we can make a conclusion that our dataset is quite reliable and both SVR and LR could be used to make reasonable predictions. They comparison result could also be seen in:

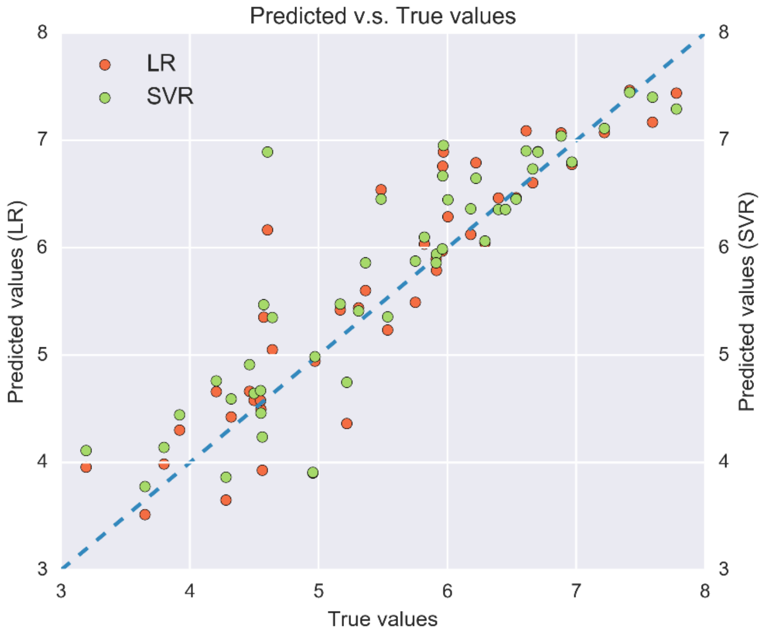


Figure 8 The comparison between predicted value of LR and SVR versus the true test value. The perfect predicted vs true value is showed in the blue dashed line. This picture verified that both algorithm could be used to make predictions.

5. Possible client of my report and suggestions

In summary, I carefully considered the dataset and model construction and pick up the most reasonable features to fit the linear regression model. In addition, I also consider the same problem by using different algorithms, after I scaled the dataset, both of them give us a similar result, which means our dataset and features selection is stable. Our final model could be summarized in the linear equation below:

$$\begin{aligned} Happiness = & -6.7844 + 0.5534 * (\text{logincome}) - 0.0358 * (\text{unemploy}) \\ & + 2.2938 * (\text{sociasupport}) + 0.0873 * (\text{publicEdu}) + 0.0104 * (\text{homicide}) \\ & + 0.0709(\text{lifeexpectancy}) - 0.0589 * (\text{logavevisitor}) + 0.4218 * (\text{logchild}) \\ & + 0.6279 * (\text{Generosity}) \end{aligned} \tag{4}$$

Our model and prediction, I believe will be useful for people, who focus their research on economy, politics, and sociology. Even for the banks, our model still has some meaning for those people. They can use our model to help other people make plans for future economic development.

Appendix:

6. All the dataset for our study in this report

Related data set: There are some available data set online that might be useful for our investigation.

1. http://worldhappiness.report/
2. https://en.wikipedia.org/wiki/Gross_National_Happiness (This is not the dataset, it is the definition of happiness)
3. http://www.fao.org/faostat/en/#data/CC
4. http://apps.who.int/gho/data/node.main.MHSUICIDE?lang=en
5. http://apps.who.int/gho/data/node.main.A1026?lang=en
6. https://www.conference-board.org/data/economydatabase/index.cfm?id=30565
7. http://data.worldbank.org/indicator/NY.GDP.PCAP.CD?view=map&year=2015
8. https://knoema.com/atlas/topics/World-Rankings