Are we happy with our life?

Author: Xiaoqian Dang
Springboard data science workshop

What makes us *Happy*?

- Money?
- Health?
- Alcohol?
- Safety?
- Education?
- Or …?

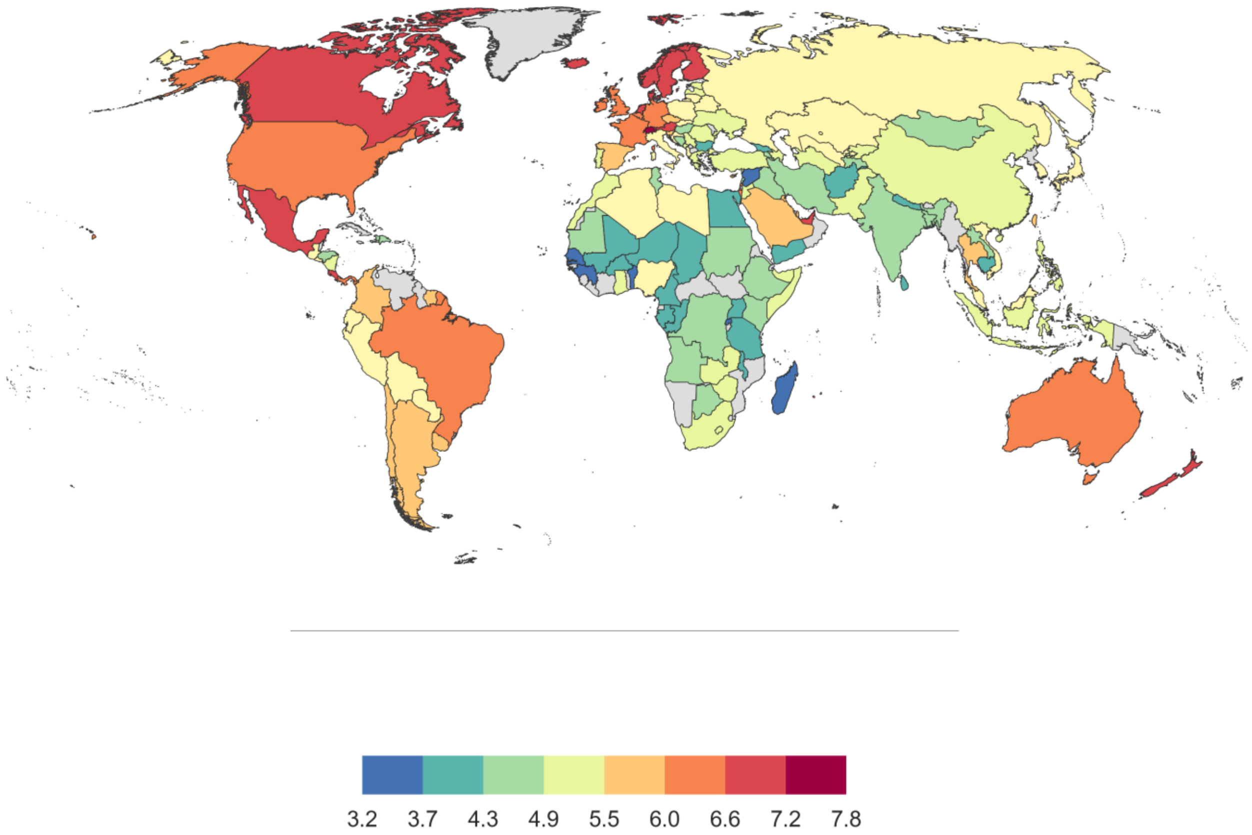
Happiness score:

“Please imagine a ladder, with steps numbered from
0 at the bottom to 10 at the top”

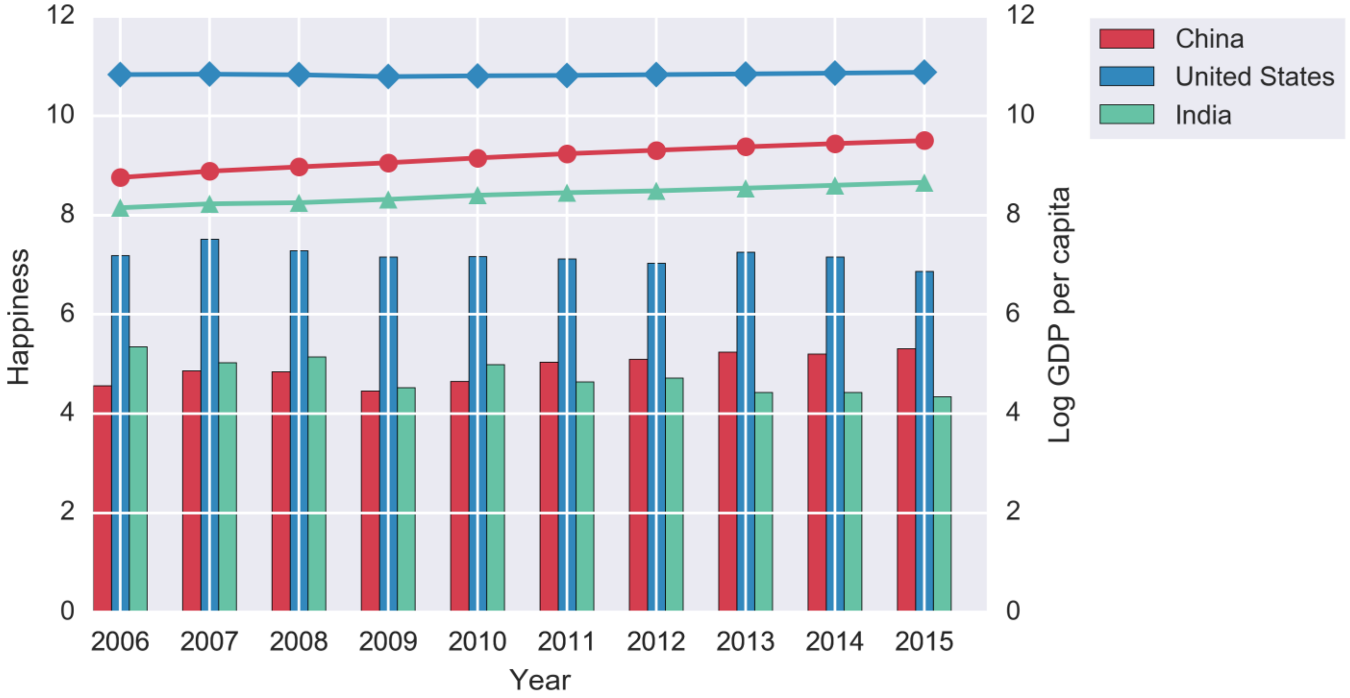
$$Happy > 5$$

$$Unhappy \leq 5$$

Happiness Score of the world in 2012

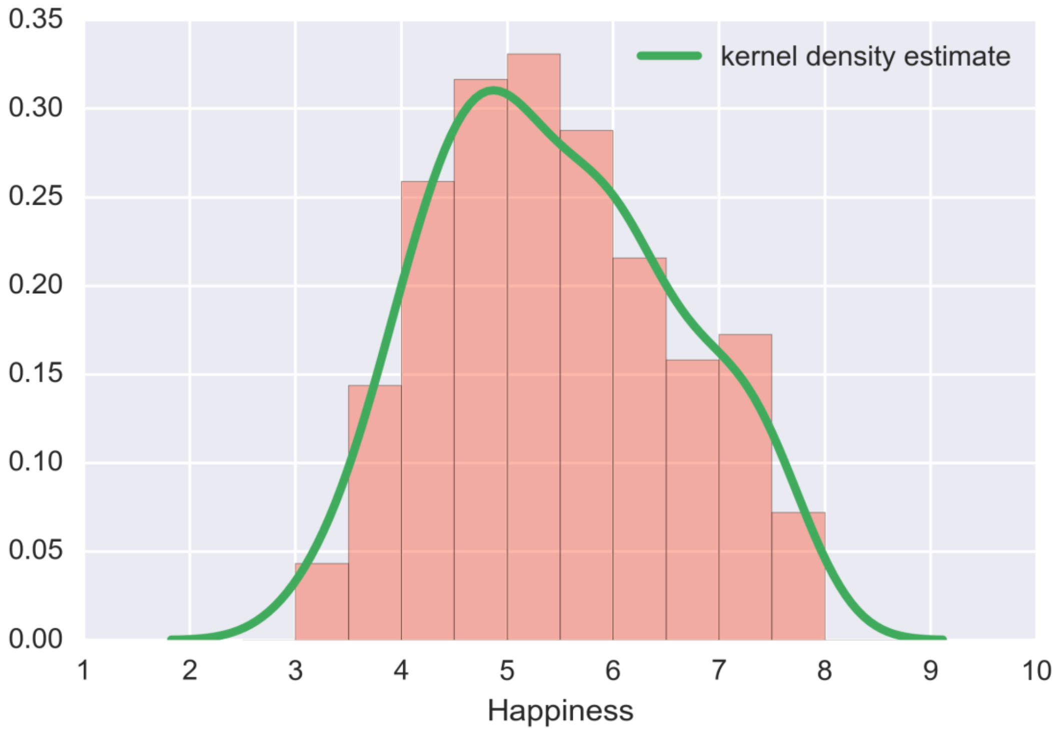


- Unevenly distribution of Happy score
- Highly related to the development level of the country
- Possible economic reason dependent!!



- Also time dependent
- Highly correlated to the value of GDP!!
- Different societies give the different Happiness score

Happiness score distribution

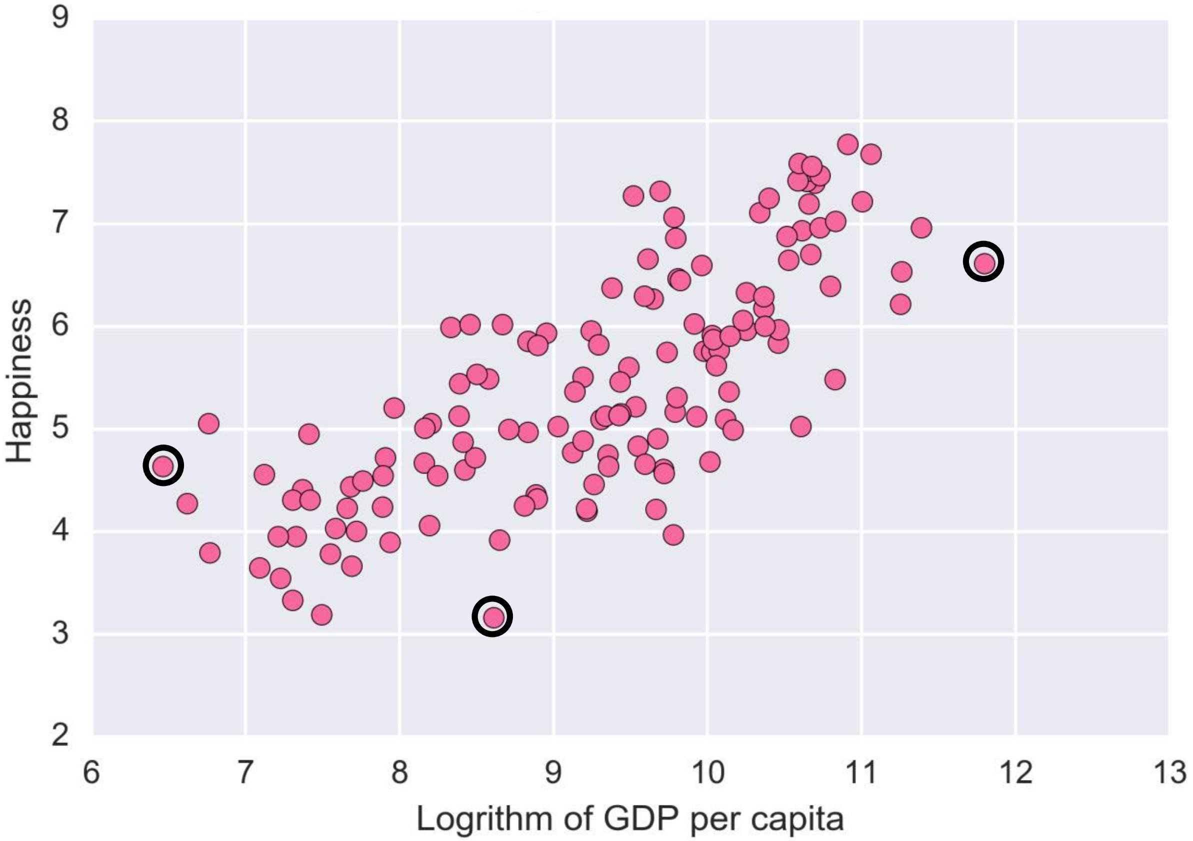


- Nearly normal distribution of Happiness score

- The average of Happiness score is little higher than the median value

- It is a sociology problem!

Happiness VS GDP



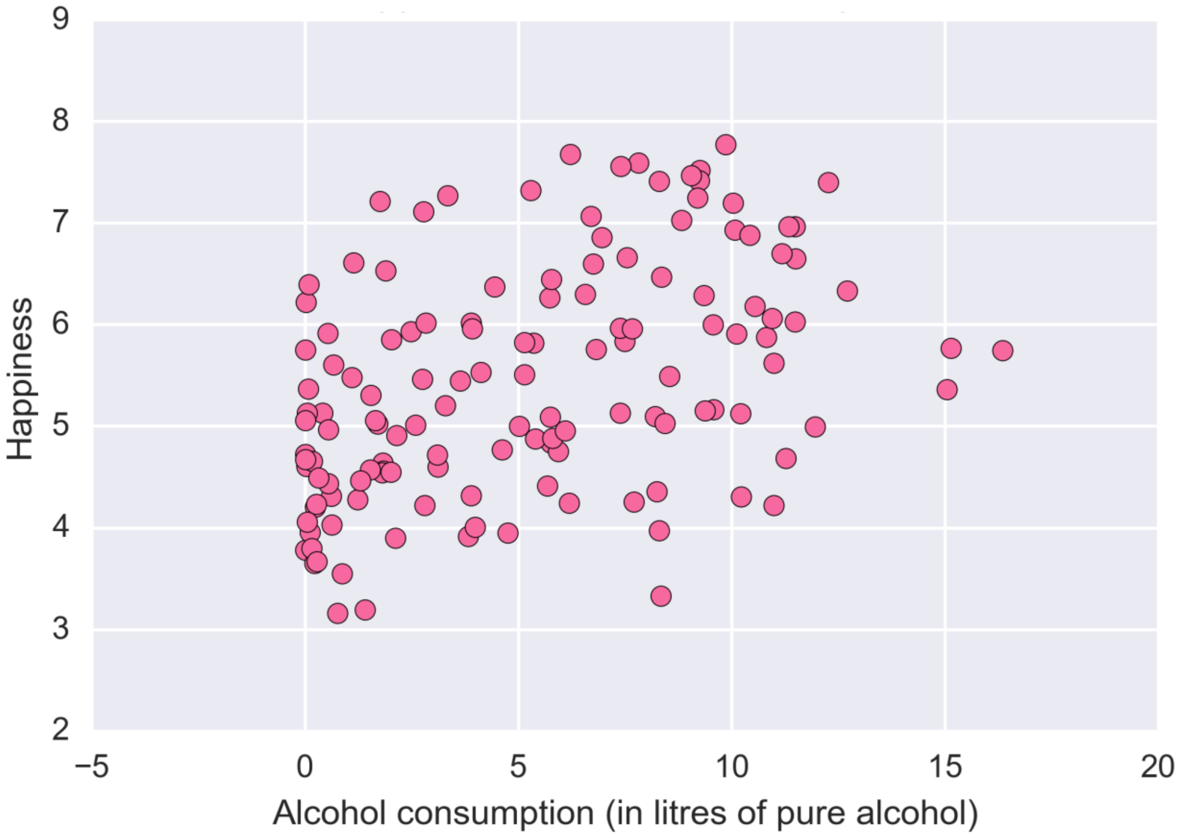
- Strongly correlated to each other!

- Three ‘abnormal’ countries indicates different relationship

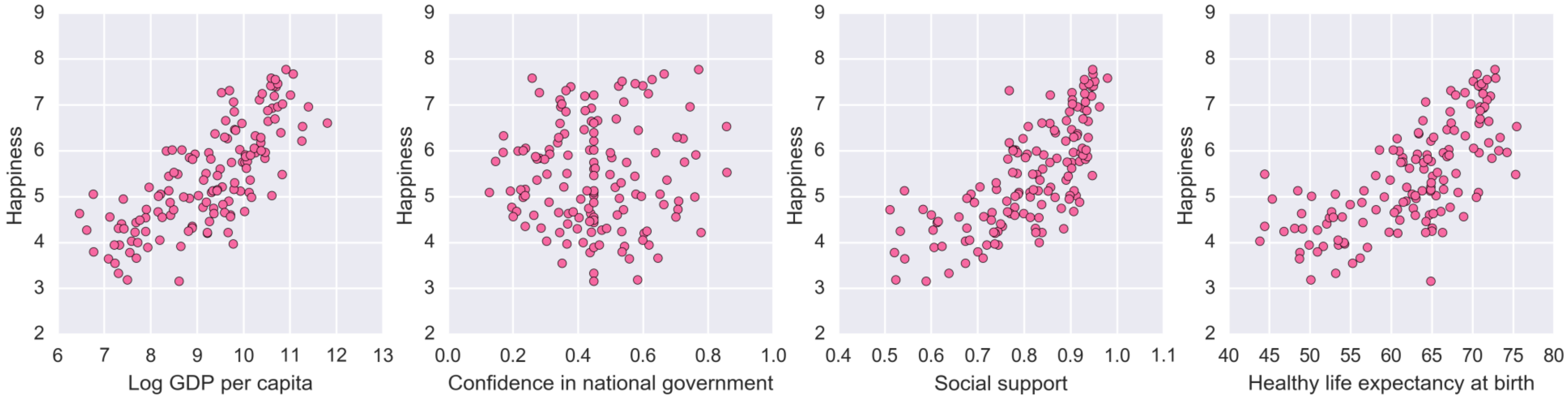
- Need more features to refine the model!!

Happiness VS Alcohol consumption

- Almost no correlation between
- Indicates Alcohol is not a good feature!!
- Need more features to refine the model!!

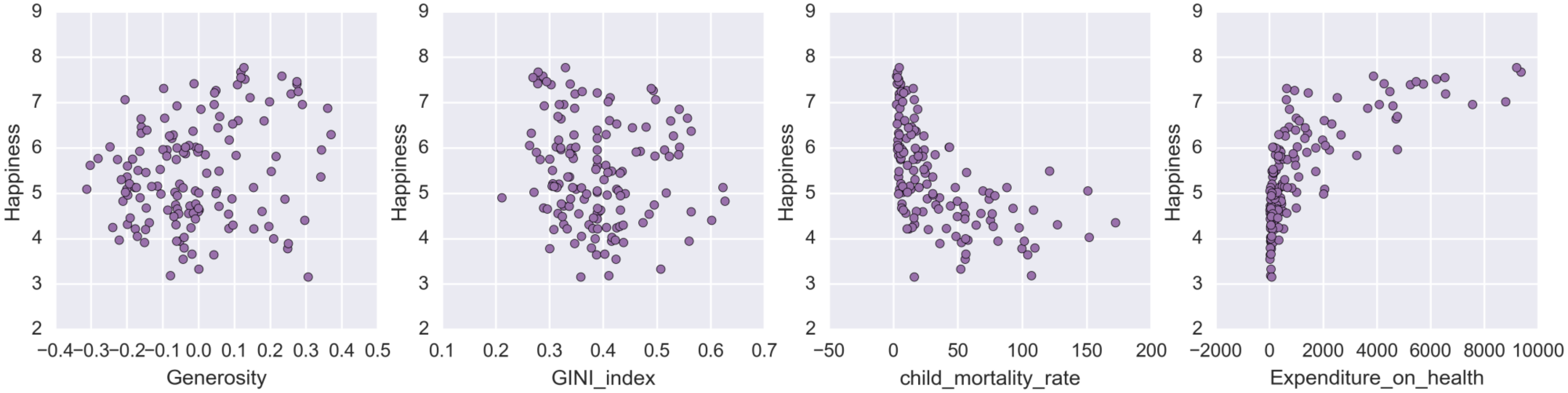


More features' correlation are on the way!!



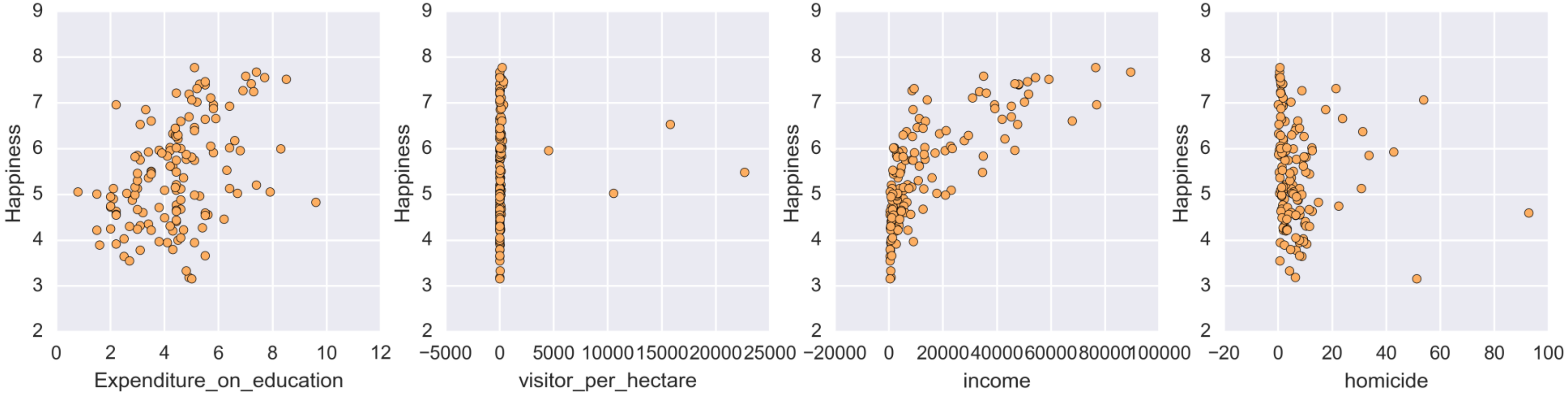
- GDP per capita
- Confidence in national government
- Social support
- Healthy life expectancy at birth

More features' correlation are on the way!!



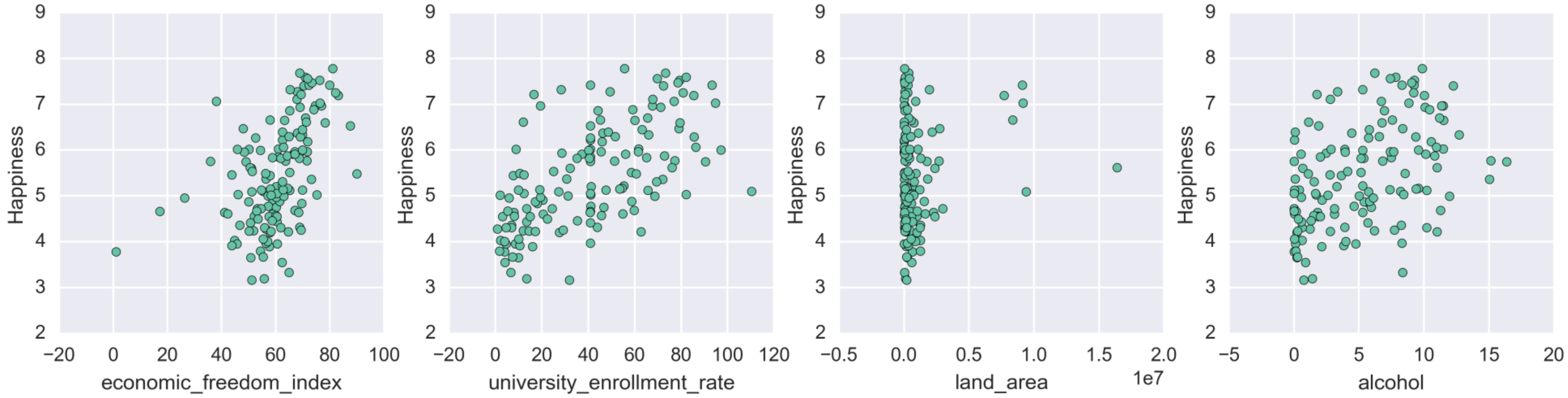
- Generosity
- GINI index
- Child mortality rate
- Expenditure on health

More features' correlation are on the way!!



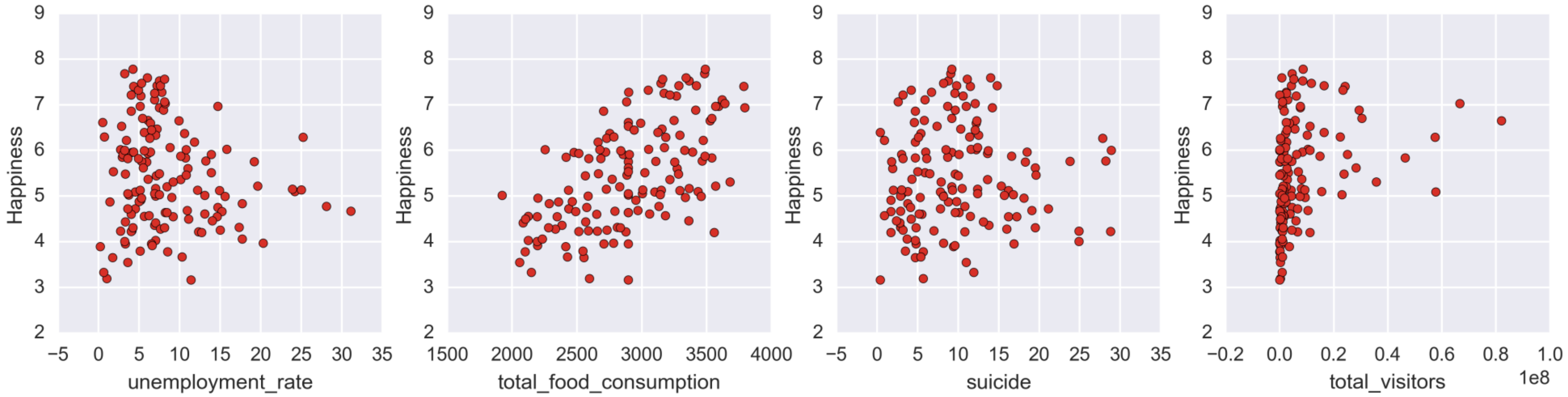
- Expenditure on education
- Visitor per hectare
- Income
- homicide

More features' correlation are on the way!!

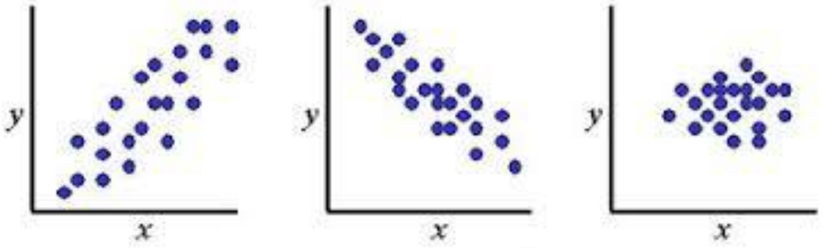


- Economic freedom index
- University enrollment rate
- Land area
- alcohol

More features' correlation are on the way!!



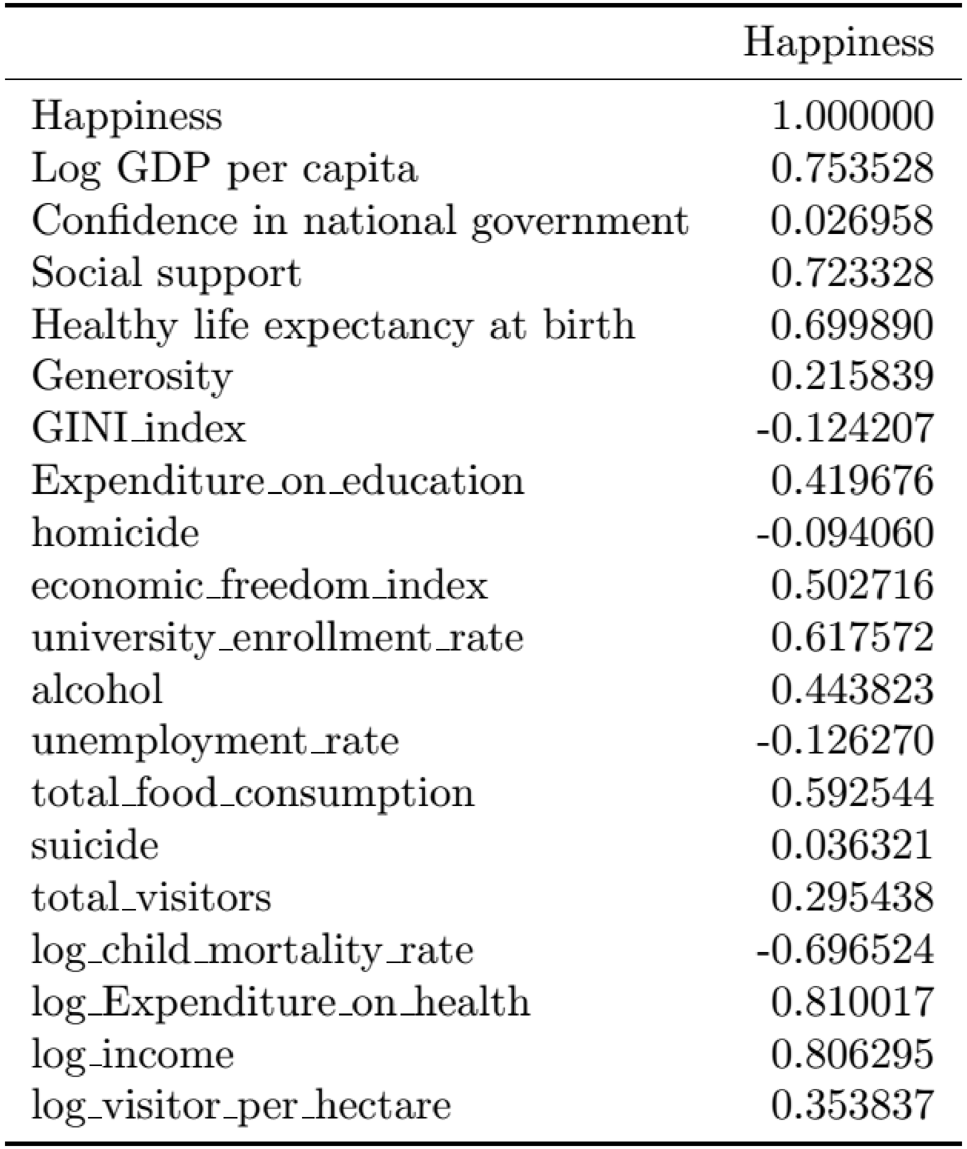
- Unemployment rate
- Total food consumption
- suicide
- Total visitors



Positive correlation Negative correlation No correlation

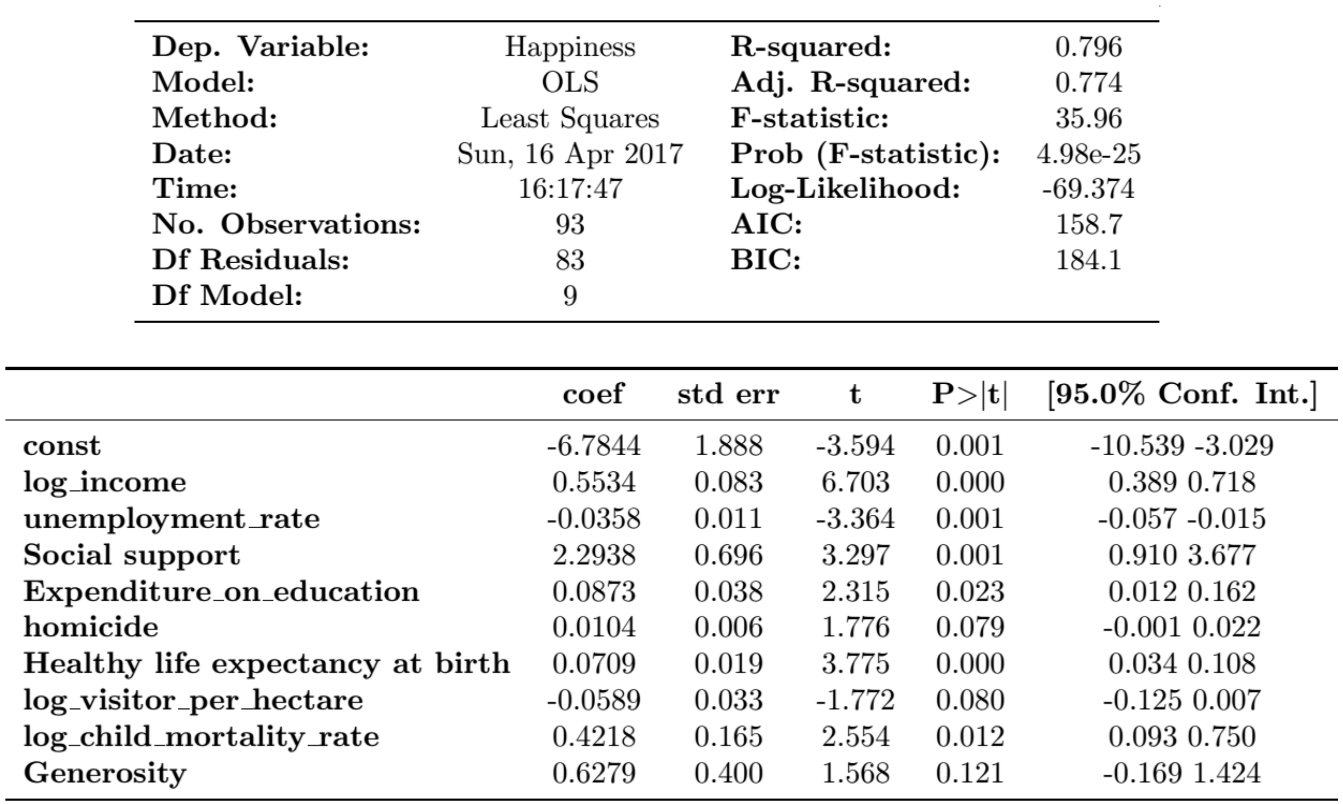
Correlation:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$



	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

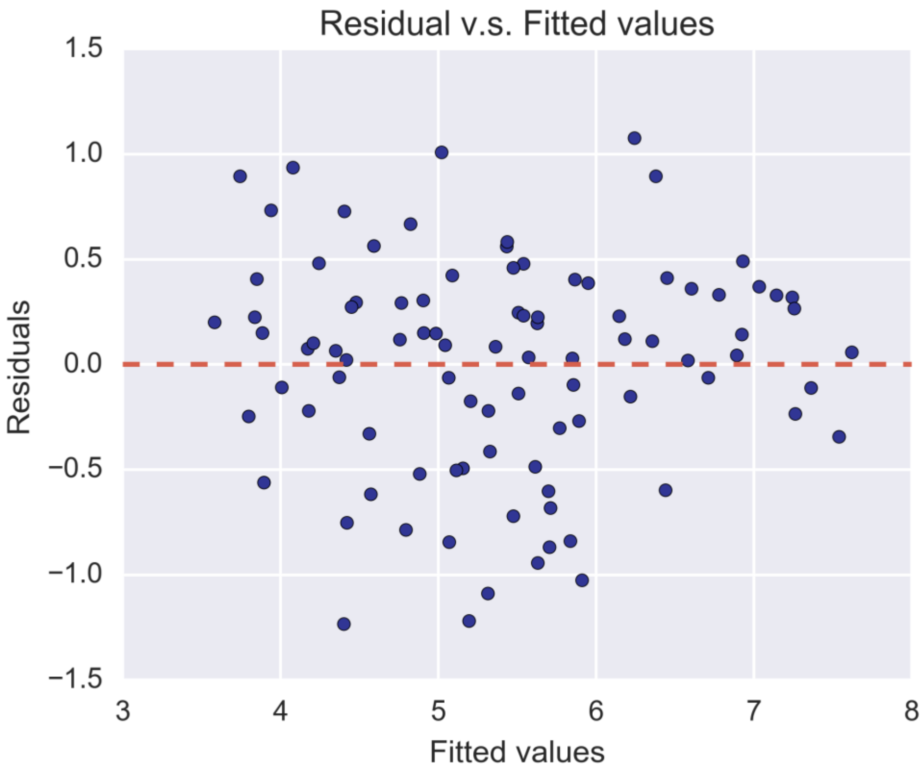
Finalized Linear Regression Model

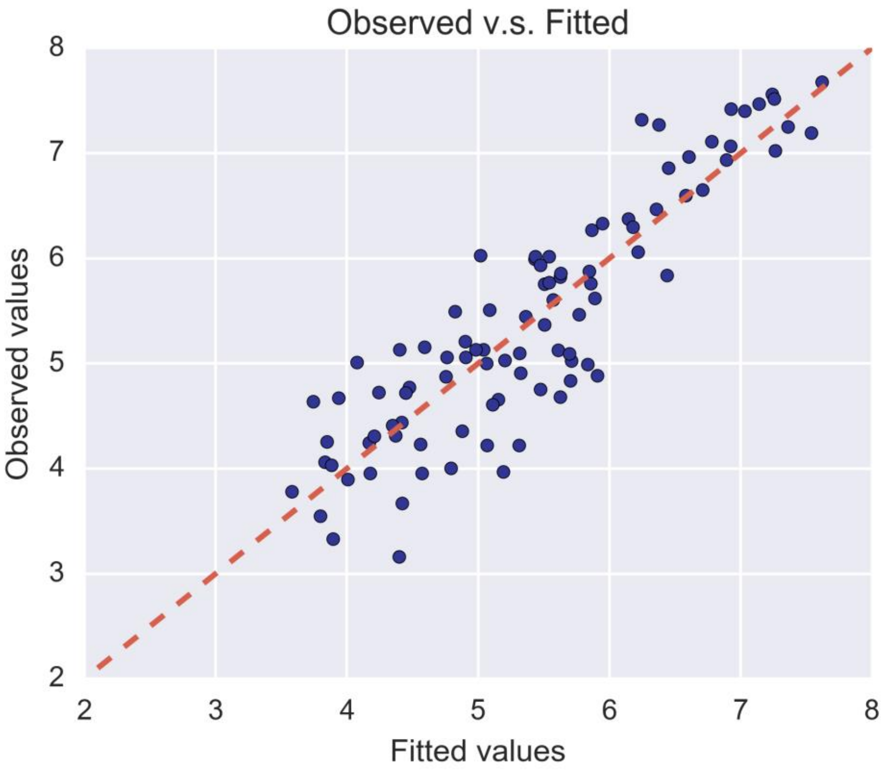


Dep. Variable:	Happiness	**R-squared:**	0.796
Model:	OLS	**Adj. R-squared:**	0.774
Method:	Least Squares	**F-statistic:**	35.96
Date:	Sun, 16 Apr 2017	**Prob (F-statistic):**	4.98e-25
Time:	16:17:47	**Log-Likelihood:**	-69.374
No. Observations:	93	**AIC:**	158.7
Df Residuals:	83	**BIC:**	184.1
Df Model: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

- A selected model with relative high R-squared value
- Features with high impact on the model prediction
- Unexpected feature selection with logarithm scale



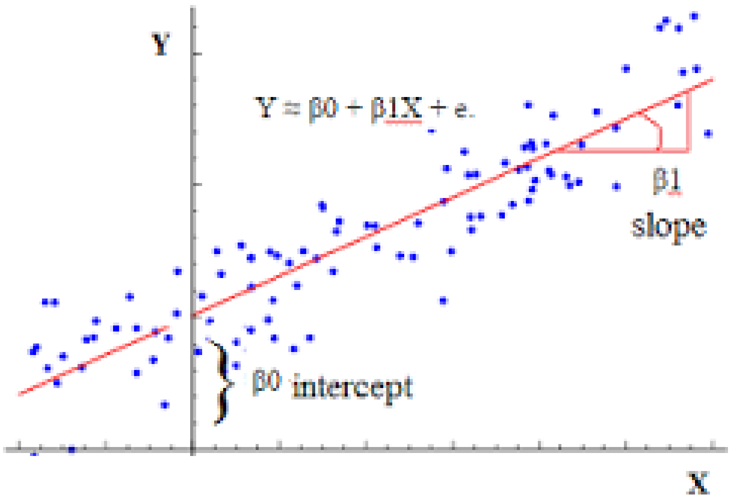


- No correlation between Residual and fitted values. A good sign to show our model prediction ability

- Almost perfect alignment between True and fitted value. But a little high variance

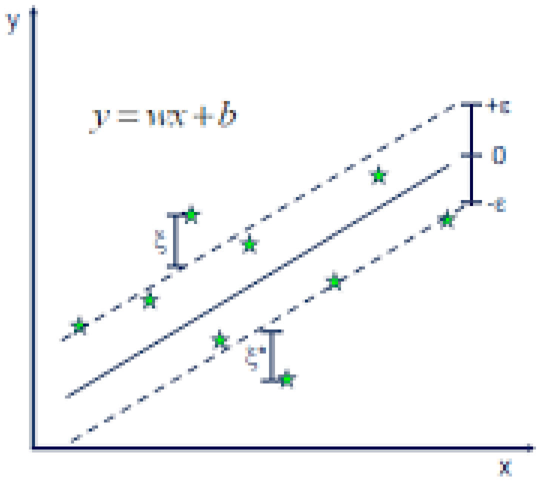
Comparison between different machine learning algorithms

Linear Regression



VS

Support vector machine



- Minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)$$

- Constraints:

$$y_i - wx_i - b \le \varepsilon + \xi_i$$

$$wx_i + b - y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

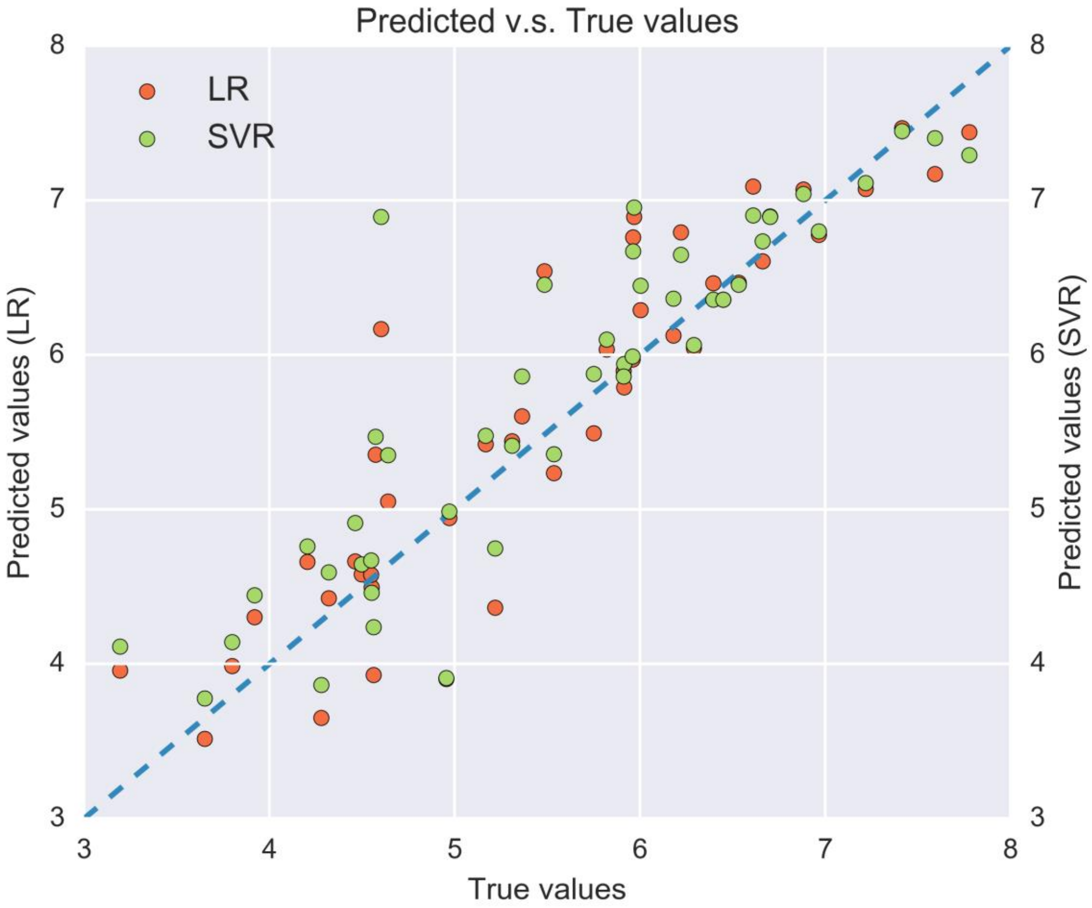
Comparison between different machine learning algorithms

The SSR score of the two different models:

$$LR : 0.24$$

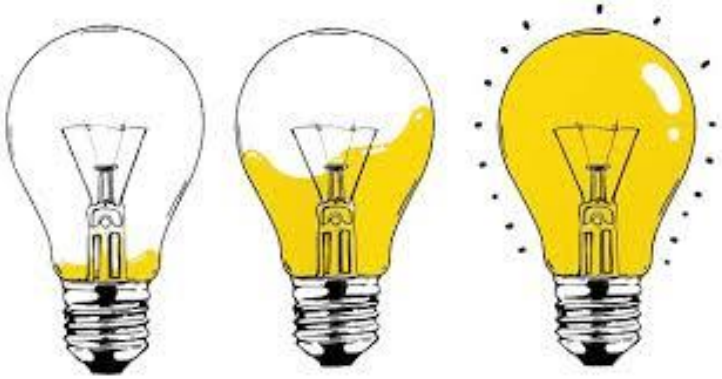
$$SVR : 0.30$$

The smaller the number, the better the fitting results.



Conclusion

- Happiness score dependents on many different features, a complicated problem
- Happiness score is strongly correlated to economic and sociology reasons
- Features selection is necessary
- Both linear regression and SVM could give us a reasonable results. The difference is not significant

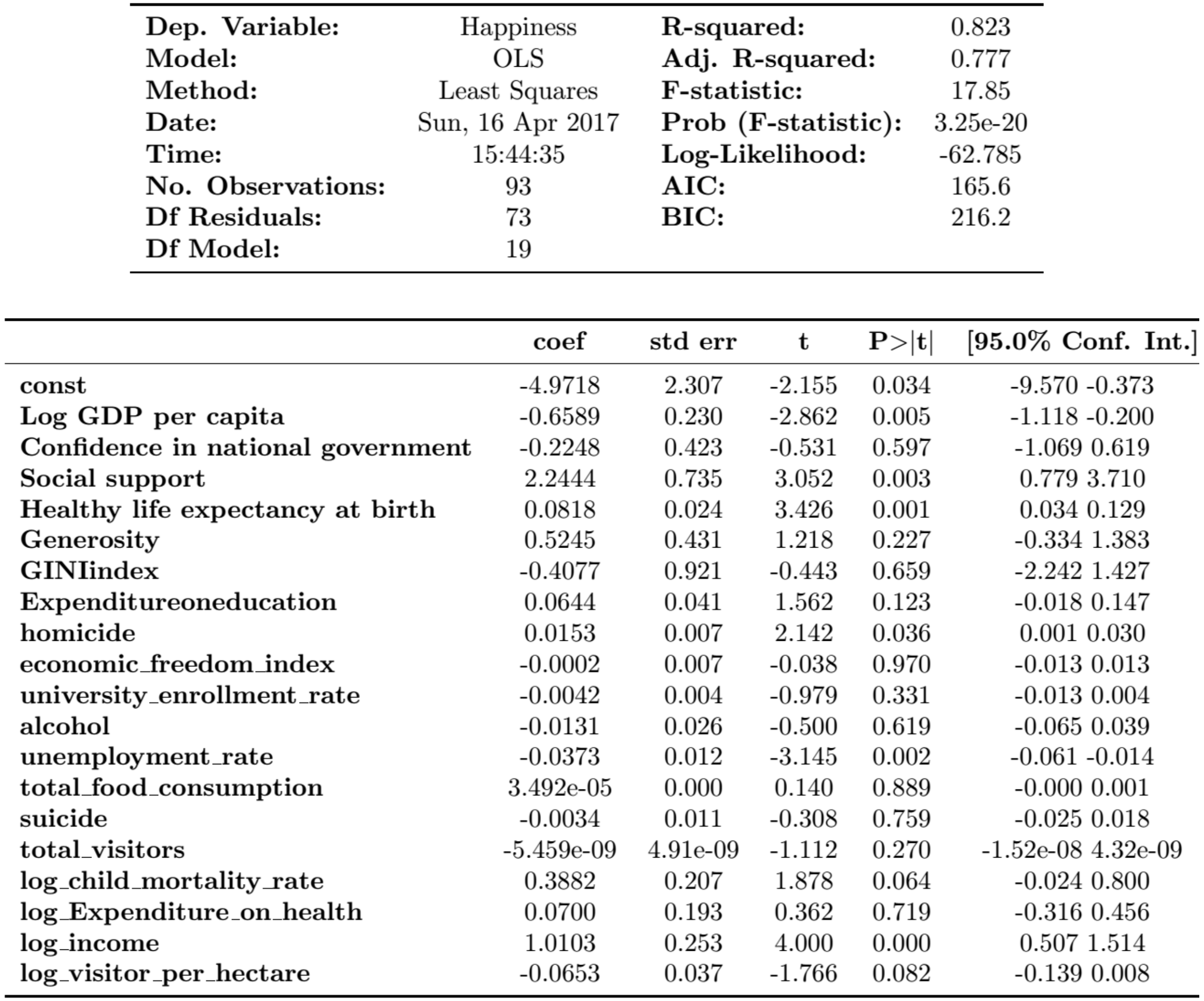


My GitHub for detailed explanation!!

References and the source of data

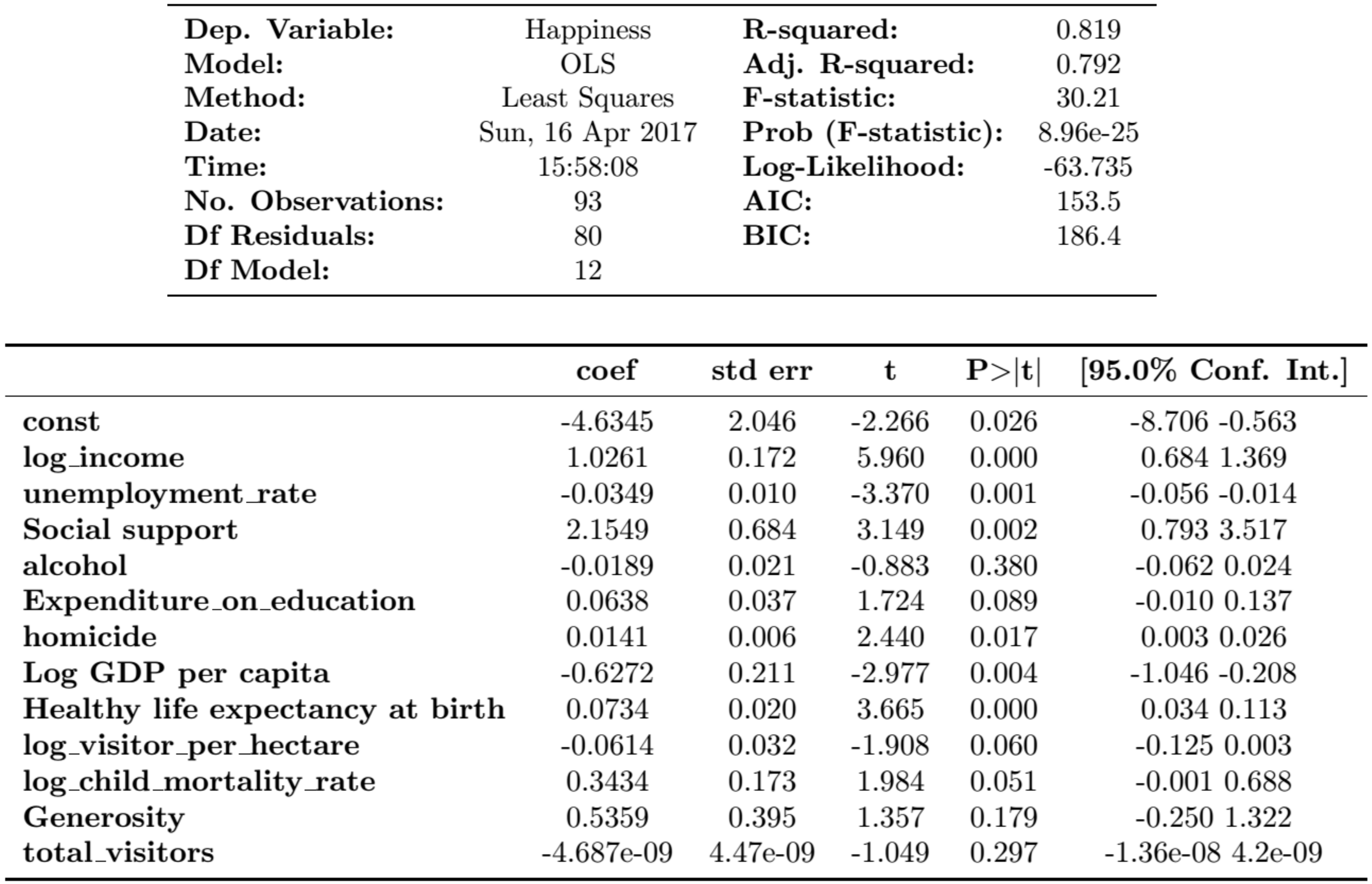
Related data set: There are some available data set online that might be useful for our investigation.

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
9. https://knoema.com/atlas/topics/Agriculture/Food-Supply-Total-Energy-kcalcapitaday/Total-food-supply
10. https://knoema.com/atlas/topics/Education/Expenditures-on-Education/Public-spending-on-education-percent-of-GDP
11. https://knoema.com/atlas/topics/Health/Health-Expenditure/Health-expenditure-percent-of-GDP
12. https://knoema.com/atlas/topics/World-Rankings/World-Rankings/Index-of-economic-freedom



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



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