Student number：20191357

An empirical study on world happiness related to geographic and social factors

Introduction

People living in different countries worldwide are having different level of satisfaction. Before 2012, GDP per capita was a key index to measure
people's satisfaction. It somehow made sense that the more money people earned, the happier they were, yet it is often not precise as factors other than money could be of importance to happiness as well. Taking Hongkong comparing with Pakistan as an example, in 2017, GDP per capita in Hong Kong was $50,000, which was 10 times of Pakistan's. However, people in Pakistan were having higher life satisfaction (Self-reported Life Satisfaction vs GDP per capita, 2017). Inspired by (Bjørnskov, Dreher and Fischer, 2006), who investigated how the government size affects national happiness, the determinants of national happiness should be explored further in terms of other social factors. Therefore, this research paper will analyses factors influencing life satisfaction of people worldwide by implementing data science approaches including visualization and regression on data collected including World Happiness Report (World Happiness Report 2017, 2017) and social progress index (E. PORTER and STERN, 2017). The domain of the studied factors will include but not limit to environment, education and healthcare. The related studies will be introduced first to derive research questions. To discover the answer of these research questions, correlation and regression study will be conducted.

Literature Review and background

Happiness has always been a heated debated and intriguing topic since it has been increasingly considered as a measure of social progress and the goal of public policy. In 2017, a full-day world happiness meeting was held in the United Arab Emirates. Before 2012, when World Happiness Report was published for the first time, the measurement of life satisfaction was often contributed by GDP per capita which is somehow ambiguous and incomprehensive. In the contrary, World Happiness Report states the satisfaction scores as sum of scores including GDP per capita, social support, healthy life expectancy, generosity, perceptions of corruption and dystopia plus residual for adjusting scores. Since published, it has been referenced and studied in variety of points of views including economics crisis (Gudmundsdottir, 2011), air pollution (Li, Folmer and Xue, 2014) and healthcare (Steptoe, 2019). Inspired by these researches, this study will focus on other factors influencing happiness, including geographics and social factors. To be specific, the following of paper will initially explore on question: which social factors affect happiness most?

Data presentation

To study the relationship between happiness and social factors. Datasets including World Happiness Report (World Happiness Report 2017, 2017) and social progress index (Porter, Stern and Green, 2017) dataset are concatenated.

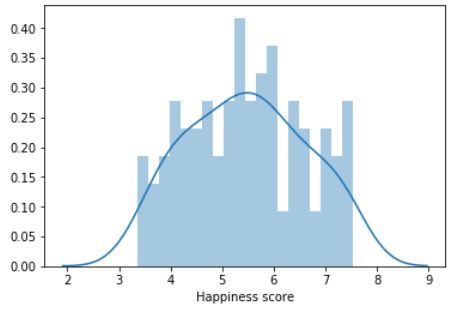

Figure 1. probability distribution of happiness score

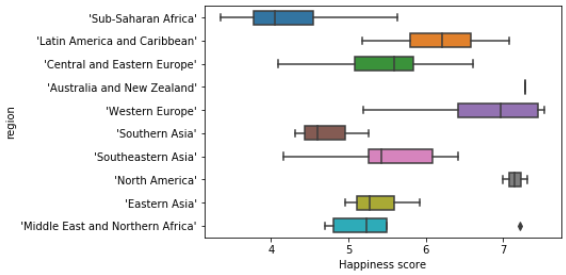

Figure 2. box plot of happiness score against region

The happiness scores are ranged from 3.4 to 7.5, with 5.47 as mean and a standard deviation of 1.14. As shown on the figure 2, the stats of happiness scores vary in different regions, where western Europe and North America having highest life satisfaction quartiles. In the contrary, sub-Sahara Africa having lowest happiness scores. Life satisfaction of other regions are ranged from 4.6 to 6.4. The only outlier is Israel located in Middle East, with a 7.3 happiness score. However, its happiness score in 2015 and 2016 are 6.8 and 7.1 respectively, proving it is not an outlier.
The descriptive data inspired another research question: Is distribution of happiness around the world geographically related?

The following sections will investigate on both social factors and geographic distribution.

Methodology

Data cleaning

As mentioned above, this study concatenates three data sources: World Happiness Report (World Happiness Report 2017, 2017) and social progress index (Porter, Stern and Green, 2017). These datasets are concatenated by the name of the country. The entries with missing data are dropped, leaving 103 entries without “nan” values.
Furthermore, the dimension of data is reduced initially by pruning columns standing for the same attribute. For instance, the high whisker and low whisker of happiness columns are dropped, leaving average happiness only, as they are highly correlated and basically illustrating the same concept in different metrics. This pruning applied to 7 columns totally.

Moreover, since the happiness score is unweighted sum of 6 scores: GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Generosity, Perceptions of corruption, Dystopia plus residual, these 6 factors are dropped as well, hauling the investigation target back to the factors related to research questions.

Happiness world map

To explore the research question 1, the visualization of data is of necessity. Plotly

(plotly.com, n.d.) is a python library that allows visualization of data in variety aspects. In this study, the geographic related components will be used to present the distribution of happiness. More specifically, a world map with happiness represented by sequential colors will be presented in this case to explore if they are clustered geographically.

Correlation matrix

As for the second research question about learning the most significant factors affecting happiness score, two methods including correlation analysis and regression analysis are adopted.

Correlation is the statical relationship between two random variables, referring to degree to which these variables are linearly related. A correlation matrix will be analyzed to find out inter-correlation of features.

Linear regression

Linear regression analysis will be conducted to study the importance of social factors using sci-kit learn package (Scikit-learn.org, 2019).

To eliminate the unfairness brought by different scales of different features, standard normalization where z = (x - mean) / standard derivation will be implemented to all features. The residuals will be analyzed to check if the linear relationship holds true.

Results

Happiness map

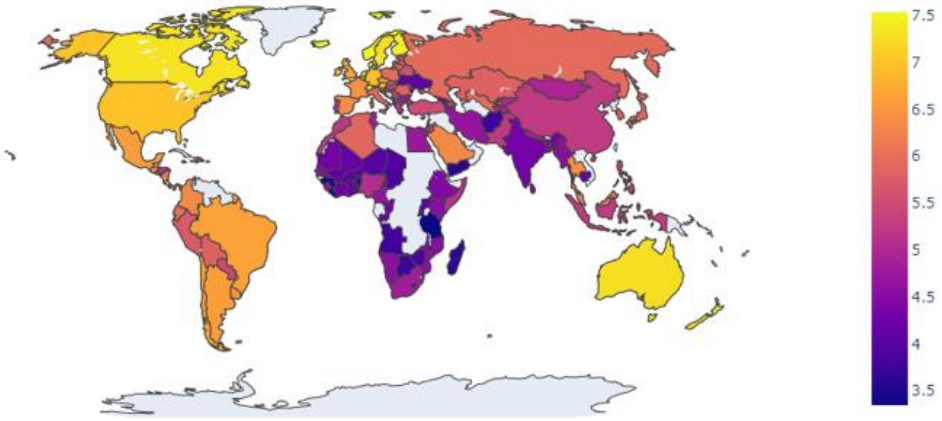

Figure 3. Happiness world map

The happiness score's geographic distribution is plotted on figure 3. The brightest areas are north America, Australia, north west Europe, while the darkest areas are centered around Africa, East Europe. It is worth mentioning that the norther part of each continent tends to have a better happiness score. Yet these observations are not sufficient to deduce happiness is geographically related. A more in-depth investigation should be conducted.

Correlation matrix

	Happiness score	Social Progress Index	Basic Human Needs	Foundations of Wellbeing	Opportunity	Nutrition and Basic Medical Care	Water and Sanitation	Shelter	Personal Safety	Access to Basic Knowledge	Access to Information and Communications	Health and Wellness	Environmental Quality	Personal Rights	Personal Freedom and Choice	Tolerance and Inclusion	Access to Advanced Education	income_inequality
Happiness score	1.00	0.34	0.30	0.34	0.34	0.23	0.28	0.30	0.27	0.25	0.24	0.42	0.32	0.28	0.33	0.29	0.28	-0.23
Social Progress Index	0.34	1.00	0.95	0.98	0.93	0.85	0.90	0.91	0.76	0.85	0.93	0.77	0.90	0.73	0.89	0.68	0.90	-0.21
Basic Human Needs	0.30	0.95	1.00	0.92	0.78	0.93	0.96	0.97	0.74	0.89	0.86	0.71	0.78	0.54	0.76	0.48	0.88	-0.24
Foundations of Wellbeing	0.34	0.98	0.92	1.00	0.88	0.84	0.87	0.90	0.71	0.86	0.92	0.83	0.93	0.68	0.85	0.64	0.85	-0.23
Opportunity	0.34	0.93	0.78	0.88	1.00	0.66	0.74	0.73	0.72	0.68	0.90	0.67	0.87	0.87	0.92	0.82	0.83	-0.14
Nutrition and Basic Medical Care	0.23	0.85	0.93	0.84	0.66	1.00	0.89	0.91	0.57	0.85	0.77	0.66	0.66	0.44	0.63	0.37	0.79	-0.19
Water and Sanitation	0.28	0.90	0.96	0.87	0.74	0.89	1.00	0.93	0.60	0.87	0.81	0.66	0.72	0.48	0.72	0.44	0.85	-0.17
Shelter	0.30	0.91	0.97	0.90	0.73	0.91	0.93	1.00	0.62	0.90	0.80	0.72	0.76	0.46	0.75	0.43	0.85	-0.24
Personal Safety	0.27	0.76	0.74	0.71	0.72	0.57	0.60	0.62	1.00	0.57	0.72	0.51	0.68	0.61	0.65	0.54	0.65	-0.32
Access to Basic Knowledge	0.25	0.85	0.89	0.86	0.68	0.85	0.87	0.90	0.57	1.00	0.75	0.56	0.68	0.41	0.71	0.38	0.82	-0.21
Access to Information and Communications	0.24	0.93	0.86	0.92	0.90	0.77	0.81	0.80	0.72	0.75	1.00	0.65	0.84	0.75	0.81	0.69	0.81	-0.22
Health and Wellness	0.42	0.77	0.71	0.83	0.67	0.66	0.66	0.72	0.51	0.56	0.65	1.00	0.78	0.56	0.62	0.56	0.58	-0.20
Environmental Quality	0.32	0.90	0.78	0.93	0.87	0.66	0.72	0.76	0.68	0.68	0.84	0.78	1.00	0.71	0.85	0.65	0.77	-0.17
Personal Rights	0.28	0.73	0.54	0.68	0.87	0.44	0.48	0.46	0.61	0.41	0.75	0.56	0.71	1.00	0.74	0.71	0.53	-0.14
Personal Freedom and Choice	0.33	0.89	0.76	0.85	0.92	0.63	0.72	0.75	0.65	0.71	0.81	0.62	0.85	0.74	1.00	0.72	0.77	-0.09
Tolerance and Inclusion	0.29	0.68	0.48	0.64	0.82	0.37	0.44	0.43	0.54	0.38	0.69	0.56	0.65	0.71	0.72	1.00	0.49	-0.04
Access to Advanced Education	0.28	0.90	0.88	0.85	0.83	0.79	0.85	0.85	0.65	0.82	0.81	0.58	0.77	0.53	0.77	0.49	1.00	-0.18
income_inequality	-0.23	-0.21	-0.24	-0.23	-0.14	-0.19	-0.17	-0.24	-0.32	-0.21	-0.22	-0.20	-0.17	-0.14	-0.09	-0.04	-0.18	1.00

Figure 4. Correlation matrix

The correlation heatmap on figure 4 is a powerful tool to visualize correlation among all the variables. As shown on Figure, apart from average income, “health and wellness”, “foundations of wellbeing” and “opportunity” are the features with highest correlation. Furthermore, other social factors also have a positive correlation between 0.23 and 0.33, showing these factors could affect happiness as well. Noticing that income inequality has a negative correlation with happiness score, indicating a higher income inequality often coming along with lower happiness score and vice versa.

Regression

	Coefficient	t-stat	R-squared	adjusted R-squared	MSE	VIF
Intercept	5.11	2.71	0.28	0.14	1.13	0.00
Social Progress Index	6.19	0.15				31244771.35
Basic Human Needs	12.4	0.32				39384566.06
Foundations of Wellbeing	0.61	0.02				16492124.73
Opportunity	26.71	0.81				26620561.72
Nutrition and Basic Medical Care	3.60	0.34				1634175.88
Water and Sanitation	3.62	0.34				6778841.84
Shelter	3.61	0.34				4429970.50
Personal Safety	3.62	0.34				2070664.69
Access to Basic Knowledge	0.68	0.07				1929112.28
Access to Information and Communications	0.63	0.07				1618396.23
Health and Wellness	0.72	0.08				1092977.64
Environmental Quality	0.65	0.07				1691352.44
Personal Rights	6.15	0.72				3494379.26
Personal Freedom and Choice	6.15	0.72				1434312.08
Tolerance and Inclusion	6.15	0.72				1666151.23
Access to Advanced Education	6.15	0.72				3562625.19
income_inequality	0.02	1.69				1.30

Table 1. Regression summary with extremely high multicollinearity

The regression was first conducted without any pre-processing of data, leaving prediction of holdout validation set with a mean squared error of 1.13 and 0.28 R-squared value. As shown on the table 1, the VIF value is extremely high, meaning that

the multicollinearity problem is severe.

Multicollinearity in linear regression study could weaken the regression and affect observational results consequently. Therefore, the feature with highest correlation will be dropped until the VIF score no longer surpass the threshold 5. Leaving 8 normalized numeric social-related features to redo the regression.

	Coefficient	t-stat	R-squared	adjusted R-squared	MSE	VIF
Intercept	5.47	53.19				0.00
Nutrition and Basic Medical Care	-0.27	-1.45				3.27
Access to Basic Knowledge	0.00	0.00				2.34
Health and Wellness	0.47	2.99				2.30
Environmental Quality	-0.06	-0.35	0.23	0.17	1.09	2.78
Personal Freedom and Choice	0.19	0.84				4.64
Tolerance and Inclusion	0.01	0.08				2.66
Access to Advanced Education	0.11	0.53				4.14
income inequality	-0.19	-1.73				1.18

Table 2. Regression summary after pruning features with VIF greater than 5

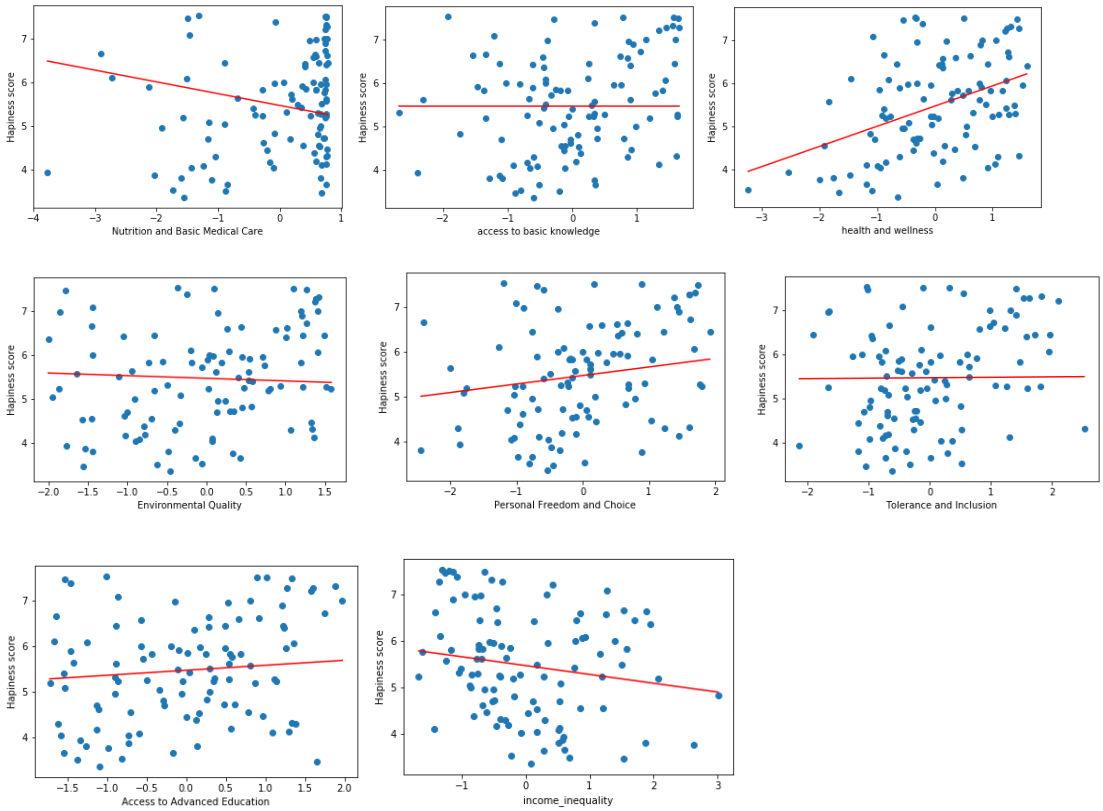

Figure 5. Visualization of Linear relationship on 8 features against happiness score

As shown on table 2, when re-implementing the linear regression after dropping features with high VIF, health and wellness is the feature with the highest positive coefficient 0.47, while nutrition and basic medical care has a greatest negative value of -0.27. income inequality is the feature with the negative coefficient -0.19, meanwhile environmental quality also affect the dependent variable negatively with the coefficient -0.06.

To the contrary, Personal freedom and choice, tolerance and inclusion and access to advanced education have a postive affect on the happiness separately, with coefficient value equal to 0.19, 0.01 and 0.11.

Therefore, the model formula can be formed as:
Happiness score= -0.27* Nutrition and basic medical care+0.47*health and wellness – 0.06environmental quality + 0.19personal freedom and choice + 0.01tolerance and inclusion + 0.11 access to advanced education – 0.19 income inequality + 5.4732

However, the R-square and adjusted R-square values are 0.23 and 0.17 respectively, along with relatively high mean square error of 1.09, showing a poor fitting of the regression.

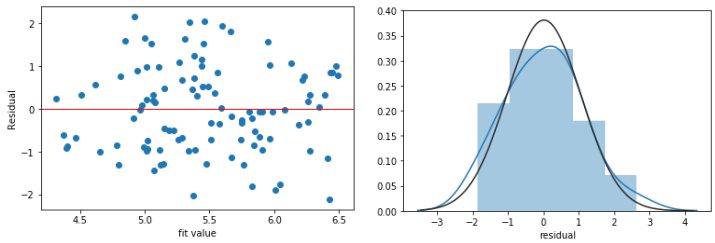

Figure 6. Residual analysis

The residual analysis is conducted as shown on figure 6, the error randomly distributed on both sides of y=0 line with no significant outliers. Furthermore, variances for residual on all fitted values are equal. The probability distribution of residual is shown on figure 5 as well, indicating the residual fitting into a normal distribution with mean at 0.

Therefore, the linear relationship is justified by the residual analysis, meaning health and wellness is the most significant factor of life satisfaction.
The same residual analysis is conducted between hapiness and income inequality, showing a valid negative linear relationship. Yet the residual analysis result for Nutrition and medical care score cannot stand out for the validity of their linear relationship.

Conclusion

In this research paper, the relationship between happiness score and social factors are studied.
The descriptive data was represented at the beginning, and the researches questions:

1. If distribution of happiness is geographically related
2. The social factors influencing life satisfaction influencing life satisfaction most.

A world map plot happiness score by sequential colors visualized the distribution of happiness score. Yet the result reveal was not sufficient to justify the hypothesis. A deep study including subdivide countries based on other geographic factors like topography, latitude and land ratio could be conducted to gain better insight on this question.

Correlation matrix and linear regression are implemented to explore relation between life satisfaction and social features. It is observed that health and wellness are the most influencing factor, while job opportunity is the second. Other factors including access to basic knowledge and foundation of wellbeing are also affecting happiness to some extents. It is worth mentioning that although multicollinearity was reduced by dropping features with high VIF, better techniques like principal component analysis could help reducing multicollinearity further.

Word count: 1677

Reference:

Axel Dreher and Justina A. V. Fischer, 2007. *The Bigger The Better? Evidence Of The Effect Of Government Size On Life Satisfaction Around The World*. Springer, pp.267-292.

Bjørnskov, C., Dreher, A. and Fischer, J.A.V. (2006). The bigger the better? Evidence of the effect of government size on life satisfaction around the world. *Public Choice*, 130(3–4), pp.267–292.

Gudmundsdottir, D.G. (2011). The Impact of Economic Crisis on Happiness. *Social Indicators Research*, 110(3), pp.1083–1101.

Li, Z., Folmer, H. and Xue, J. (2014). To what extent does air pollution affect happiness? The case of the Jinchuan mining area, China. *Ecological Economics*, 99, pp.88–99.

Matplotlib.org. (2012). *Matplotlib: Python plotting — Matplotlib 3.1.1 documentation*. [online] Available at: https://matplotlib.org/.

Our World in Data. (2017). *GDP per capita vs Self-reported Life Satisfaction*. [online] Available at: https://ourworldindata.org/grapher/gdp-vs-happiness.

plotly.com. (n.d.). *Modern Analytic Apps for the Enterprise - Plotly*. [online] Available at: https://plotly.com/.

Porter, M., Stern, S. and Green, M. (2017). *SOCIAL PROGRESS INDEX 2017 SOCIAL PROGRESS INDEX 2017*. [online] Available at: https://www2.deloitte.com/content/dam/Deloitte/co/Documents/about-deloitte/Social-Progress-Index-2017.pdf.

Pydata.org. (2012). *seaborn: statistical data visualization — seaborn 0.9.0 documentation*. [online] Available at: https://seaborn.pydata.org/.

Scikit-learn.org. (2019). *scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation*. [online] Available at: https://scikit-learn.org/stable/.

Steptoe, A. (2019). Happiness and Health. *Annual Review of Public Health*, 40(1), pp.339–359.

Worldhappiness.report. (2017). *Home*. [online] Available at: https://worldhappin Reference list