Happiness Scores Analysis Report

Lenong Xu(✉)

The University of Manchester, Manchester, England
503352480@qq.com

Abstract. Happiness scores are often used to inform policy decisions. This report is dedicated to analyzing happiness scores to support better government decision-making. In the exploratory data stage, statistical analysis, comparative distributions, regression analysis, and correlation analysis are carried out using Excel and Stata. Predictive models of regression and neural networks were built after cluster analysis using SAS Enterprise Guide and SAS Enterprise Miner Workstation. Average Squared Error (ASE) was a used to select the better regression prediction model. The impact of the COVID-19 and the public policies after the pandemic are briefly discussed.

Keywords: Happiness scores · Cluster analysis · Predictive models · COVID-19

1 Introduction

Happiness is seen as a fundamental human goal. The United Nations released the first World Happiness Report in 2012 and declared 20 March as International Happiness Day [1]. However, the COVID-19, which is the most serious health threat of the century, emerged in 2019. As the number of coronavirus illnesses and even deaths continued to grow, the happiness of people is greatly affected [2].

The aim of this report is to analyze happiness scores and the factors that influence them to predict well-being scores and to inform public policymaking on happiness. These datasets are mainly from the Gallup World Poll and the ladder scores in the datasets are exactly the happiness scores.

The challenge of this project is the quality problem of the datasets with outliers, missing data and redundant data that can affect the efficiency and reliability of the data analysis. Therefore, based on the understanding of the data, data preparation must be carried out before the data exploration part. Then after using clustering analysis to cluster countries into suitable clusters, the main features of each segment were obtained. Additionally, due to the complexity of the influences on happiness scores, there is a need to find the most accurate prediction model to predict happiness scores. To this end, regression and neural network models were built, and ASE was used to evaluate the performance. The data analysis is then used as the basis for a brief discussion of the impact of the epidemic and post-COVID-19 public policy. Finally, the conclusions and limitations of the project are integrated.

G. Ali et al. (Eds.): ISEMSS 2022, ASSEHR 687, pp. 2892–2901, 2022.
https://doi.org/10.2991/978-2-494069-31-2_340

2 Exploratory Data Analysis

2.1 Data Understanding

In this report, the “Happiness2021” dataset is referred to as dataset 1, “Happiness-byYears” dataset as dataset 2.

2.1.1 Data Variables

Dataset 1 describes information on the regional indicator for 149 countries as well as the ladder score, logged GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption, the last six of these are variables that affect the ladder score, i.e., the happiness score.

Dataset 2 contains happiness scores for 166 countries in different years (1 ~ 15 years), the same six variables as in dataset 1 and the addition of positive affect and negative affect variables.

2.1.2 Data Quality

The data in dataset 1 is detailed, well organized, with completeness, conformity and consistency. The years for which data are available vary across countries in dataset 2. The report only wants to analyze the correlation between happiness scores and positive/negative impacts in 2020 for each country, so there is a considerable amount of redundant data in the dataset 2. Also, the presence of a large amount of missing data can be clearly observed.

2.2 Data Preparation

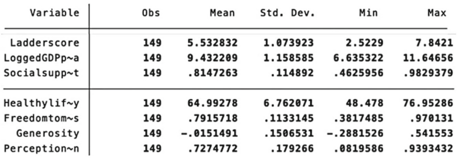
Data set 1 contains 1 outlier which may affect the process of estimating the statistics and lead to over or under valuation. However, it is not intended to deal with this outlier in this project, but rather as a special sample for analysis.

There was redundant data in dataset 2, which needed to be filtered first for the year 2020 and to remove variables other than positive affect and negative affect. There were 94 items of filtered data, of which one had missing data in the positive affect and negative affect sections. The presence of missing values means that less data is available for analysis, leading to biased results and reduced data efficiency, thus affecting the reliability of the results [3]. As the missing data represent a relatively small proportion of the data, regression analysis in Stata will be used to directly analyze the relationship between the happiness scores and the two variables.

2.3 Data Exploration

2.3.1 Distribution for Regional Groups

The 149 countries in dataset 1 are in the 10 regional groups. Observations show that Western Europe and North America and ANZ have the highest happiness scores, while Western Europe and South Africa have an upper box shape, indicating that most of the



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 Ladderscore |        149    5.532832    1.073923     2.5229     7.8421
LoggedGDPp~a |        149    9.432209    1.158585   6.635322   11.64656
Socialsupp~t |        149    .8147263     .114892   .4625956   .9829379
-------------+---------------------------------------------------------
Healthylif~y |        149    64.99278    6.762071     48.478   76.95286
Freedomtom~s |        149    .7915718    .1133145   .3817485    .970131
  Generosity |        149   -.0151491    .1506531  -.2881526    .541553
Perception~n |        149    .7274772     .179266   .0819586   .9393432
```

Fig. 1. Stata's summarize table

data are more skewed towards high happiness scores. Middle East and North Africa and Sub-Saharan Africa are far apart from the upper and lower edges, indicating a wider distribution of data.

It is worth noting that one country in Latin America and the Caribbean, Haiti, has an unusually low happiness score of 3.615. Haiti is an extremely backward country with economic, political and social difficulties and a number of natural disasters that have plunged the country into chronic poverty and other serious problems. The mortality rate in Haiti is high, rural Haitians still believe that the government has failed to bring them security, health care, clean water and a viable transportation system and that most of the population believes that officialdom is corrupt [4]. These reasons may all contribute to the low happiness scores of the Haitian people.

The analyse of the mean and median happiness scores for each region shows that North America and ANZ has the highest happiness scores with a mean of 7.116 and a median of 7.103. Western Europe ranks 2nd with a mean and median of 6.416 and 6.282 respectively. South Asia has the lowest happiness score, with a mean and median of only 4.521 and 4.721 respectively.

2.3.2 Effect of Factors on Happiness Scores

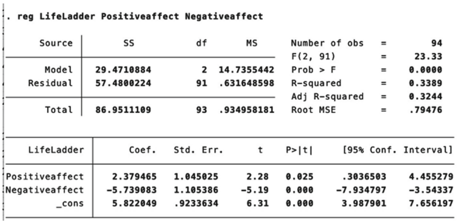
Using Stata's summarize yields the following table, which gives detailed information on the happiness scores and variables (Fig. 1).

An analysis of the correlation coefficients between each variable and the happiness score was produced. The correlation coefficients of generosity and perceptions of corruption are less than 0, i.e. they are negatively correlated with happiness scores. The absolute value of the correlation coefficient between generosity and happiness score is less than 0.4, so it is a low correlation. The absolute values of the correlation coefficients for perceptions of corruption and freedom to make life choices are between 0.4 and 0.7, so they are moderately correlated. The absolute values of the correlation coefficients for the remaining variables are greater than 0.7, so they are highly correlated.

2.3.3 The Impact of Positive and Negative Emotions

To exclude the interference of missing values, regression analysis was conducted using Stata. Figure 2 shows that the p-value is less than 0.05, so both factors have a significant effect on the happiness score.

The coefficient of positive emotion is positive, so it is a positive influence, i.e., the more positive the emotion, the higher the happiness score. At the same time, the coefficient for negative emotions is negative and the absolute value of the coefficient



```
. reg LifeLadder Positiveaffect Negativeaffect

      Source |       SS           df       MS      Number of obs   =        94
-------------+----------------------------------   F(2, 91)        =     23.33
       Model |  29.4710884         2  14.7355442   Prob > F        =    0.0000
    Residual |  57.4800224        91  .631648598   R-squared       =    0.3389
-------------+----------------------------------   Adj R-squared   =    0.3244
       Total |  86.9511109        93  .934958181   Root MSE        =    .79476

------------------------------------------------------------------------------
  LifeLadder |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Positiveaffect |   2.379465   1.045025     2.28   0.025     .3036503    4.455279
Negativeaffect |  -5.739083   1.105386    -5.19   0.000    -7.934797    -3.54337
       _cons |   5.822049   .9233634     6.31   0.000     3.987901    7.656197
------------------------------------------------------------------------------
```

Fig. 2. Regression in Stata

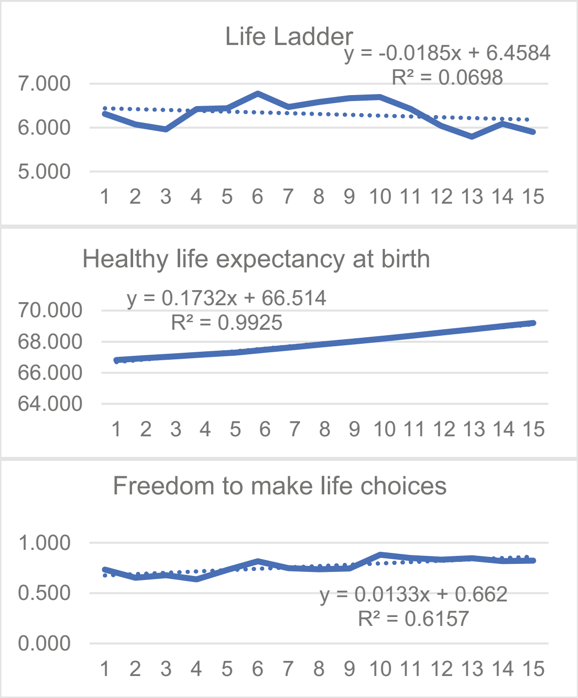
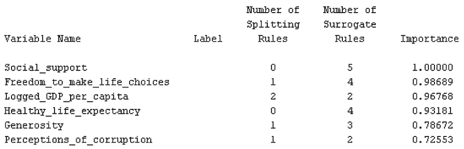


Fig. 3. Happiness scores in Argentina

is even more than twice that of positive emotions, which means that negative emotions have a very negative impact on happiness score.

2.3.4 Analysis of Argentina in the Last 15 Years

Argentina's data from 2006 to 2020 is complete, organized and without missing data. Argentina's data was filtered in dataset 2 to produce a line graph with trend line and equation for happiness scores as well as factors over a 15-year period (Fig. 3). Clearly, the ladder score shows an increasing trend from 2006 to 2015 and a decreasing trend after 6.697 in 2015.The values for Healthy life expectancy in birth and Freedom to make life choices have been increasing year on year.



```
                                        Number of    Number of
                                        Splitting    Surrogate
Variable Name                  Label      Rules        Rules      Importance

Social_support                             0            5          1.00000
Freedom_to_make_life_choices               1            4          0.98689
Logged_GDP_per_capita                      2            2          0.96768
Healthy_life_expectancy                    0            4          0.93181
Generosity                                 1            3          0.78672
Perceptions_of_corruption                  1            2          0.72553
```

Fig. 4. Importance of variables

Analyze the correlation, the healthy life expectancy in birth and freedom to make life choices are highly correlated, while generosity and negative influence are moderately correlated as they tend to decrease over time.

3 Clustering Analysis

3.1 Cluster Countries

Use SAS Enterprise Guide to convert the format of the excel file and import it into SAS Enterprise Miner Workstation. Run the cluster node to cluster countries into 4 segments and rank the segment frequencies from segment id 1 to 4 from largest to smallest as 21, 76, 33, 19 respectively.

As can be seen from Fig. 4, social support is the most important of all the variables at 1. Freedom to make life choices, logged GDP per capita and healthy life expectancy follow in decreasing order of importance, but they are all above 0.9, i.e., they are also very important. The rest of the variables have a significance between 0.7 and 0.8.

3.2 Key Features of Country Clusters

Running the segment profile node yields the following profiles and the value ranking of each factor on the happiness score.

3.2.1 Segment 1

In Segment 1, perceptions of corruption are distinctly below average and clearly above average in terms of logged GDP per capita, healthy life expectancy, freedom to make life choices and social support. The social environment in this part of the country is excellent, with people living in abundance, health and without concern for corruption. The perceptions of corruption, which are significantly below average, have the greatest impact on the happiness scores, while generosity has the least impact.

3.2.2 Segment 2

Segment 2 has the highest segment frequency, which is probably why the levels are all close to the average. Logged GDP per capita is the most important factor in terms of happiness, followed by generosity, healthy life expectancy and social support in decreasing order of importance.

3.2.3 Segment 3

In Segment 3, healthy life expectancy, logged GDP per capita and social support are all considerably below average. Freedom to make life choices is also below average. These countries are in the lower middle of the range in terms of health care and economic development. The three factors that most affect the happiness of people in these countries are the same three that are significantly below average.

3.2.4 Segment 4

In Segment 4, generosity is significantly above average, but below average in terms of logged GDP per capita, healthy life expectancy and social support. These countries are not great in their development but are still generous even with poor living standards. At the same time, generosity has a disproportionate impact on happiness in these countries.

4 Predictive Modelling

This module will develop prediction models through regression and neural networks. The performance of the models will be compared based on ASE and the better performing prediction model will be selected.

4.1 Building Predictive Models

After importing dataset 1 in SAS Enterprise Miner Workstation, insert the data partition node. Select the data partition node and set 70% of the input dataset as the training set and 30% as the validation set. The training set is used for initial model fitting and the validation set is used to compare and evaluate the performance of the prediction model. Then, the regression and neural network models were built.

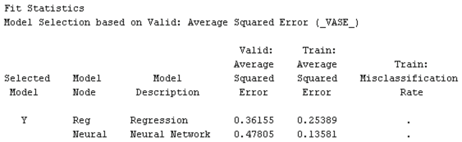
4.1.1 Regression Model

Firstly, build the linear regression model. It is a method of predicting a target variable by fitting the best linear relationship between the dependent and independent variables [5]. Insert the regression node, change the regression type to linear regression and the selection criterion to validation error and run the node.

4.1.2 Neural Network Model

The neural network model is then built, inspired by the way biological neural networks in the human brain process information. Using the Multilayer perceptron, such networks consist of multiple layers of computational units, usually interconnected in a feed-forward manner. Each neuron in one layer has a direct connection to a neuron in the next layer [6].

Insert the neural network node, change the Model Selection Criterion to Average Error and run it.



```
Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

                                          Valid:     Train:
                                         Average    Average           Train:
Selected    Model         Model          Squared    Squared    Misclassification
 Model      Node       Description        Error      Error            Rate

   Y        Reg       Regression         0.36155    0.25389             .
            Neural    Neural Network     0.47805    0.13581             .
```

Fig. 5. ASE of the prediction models

4.2 Evaluating Predictive Models

Connect models to the model comparison node, change selection statistic to Average Squared Error and select table to validation. Observing the table of ASE values (Fig. 5), the regression prediction model has the smaller ASE of approximately 0.36, meaning that it has better performance and can predict more accurate results.

5 Data Analytics to Inform Public Policy

5.1 Effects of COVID-19 on Happiness, Positive/Negative Emotions

The COVID-19 epidemic provided ample cause for negative emotions. Positive emotion also declined sharply during the March 2020 virus outbreak [7]. There is fear and anxiety in anticipation of the possibility of contracting the virus, sadness over reduced social connections, and anger and frustration over the loss of jobs, income and freedom [8]. More seriously, the pandemic brings with it not only a decline in living standards and economic status, but also the threat of death. As of 8 December 2021, there are already more than 267 million cases of COVID-19 worldwide, and nearly 5.3 million deaths as a result [9].

The negative impact of the dramatic increase in negative emotions and the decrease in positive emotions on the happiness of people is extremely significant, and the deterioration of mental health follows. According to the survey, 13.6% of people in the US exhibit serious mental illness, compared to 3.9% in 2018 [10]. Similarly, within the first month of the Australian COVID-19 restrictions, approximately one quarter of adult respondents suffered from depression or anxiety symptoms, which was significantly higher than the 2014 systematic evaluation (subthreshold anxiety, 4.4%) [11].

5.2 Post-cOVID-19 Public Policies

First of all, the negative impact caused by COVID-19 has to be addressed in its own right, which is why preventive and control measures against the epidemic are essential. Countries must strengthen their medical and epidemiological policies to protect people's lives and health. For example, with the rapid spread of the Omicron variant in the UK, the Prime Minister confirmed on 8 December that England would move to Plan B to strengthen its vaccination measures [12].

In addition, the financial worry that comes with being unable to work has a negative impact on happiness. The World Happiness Report states that unemployment during the pandemic was associated with a 12% decrease in life satisfaction and a 9% increase in negative impacts [13]. In South Africa, where policies are strict, the immediate consequences of the pandemic could lead to an additional 3 to 7 million people being unemployed in 2020, thereby increasing the unemployment rate to around 50% [14]. Therefore, the policy makers must consider how to improve the employment situation, such as providing employment and improving the remote working environment. A proper relaxation of the rules and giving people more freedom of choice in their lives might also greatly enhance happiness.

It is worth noting that China has initiated a mental health support system, in addition to physical support, to cope with the psychological stress that prevailed during the epidemic and its aftermath [15].Increased information sharing and interaction between the government and the public, transparency in news about the epidemic, increased publicity on epidemic prevention and the provision of mental health assessments and psychological counselling can enhance trust in the government and thus create a positive social atmosphere [16].

6 Conclusions

In conclusion, this report examines the happiness index and six important influencing factors. Data such as the Gallup World Poll happiness index and the ASE were used as indicators to assess the performance of the prediction models and to select the better performing linear regression prediction model. The study found that the epidemic also had a significant negative impact on people's well-being, so the data-based analysis also informs public policy making in the aftermath of the epidemic.

7 Limitations

Regarding data sources, the dataset is primarily derived from the Gallup World Poll, which samples 1,000 completed questionnaires in a country [17]. Although some major cities collect larger samples, the sample size is still too small for all population, which can lead to differences between the data analysis and the actual situation.

For the dataset, the six variables utilized are assumed to be influences on happiness scores and are used as the basis for analyzing the data and predicting the happiness scores of the imagined countries. However, the World Happiness Report contains more variables than those analyzed in this report [18]. Further research is needed to consider factors that have not been analyzed as well.

Acknowledgments. First of all, I would like to thank my supervisor, Professor Yu-wang Chen. He accompanied me through the whole process of completing my thesis and gave me a lot of guidance and inspiration. Secondly, I would like to thank my university, the University of Manchester, for giving me excellent conditions and motivation to study at this great university. Finally I would like to thank my family and friends who have supported me through the process of completing my report.

References

1. Oberoi, P., Chopra, S. and Seth, Y. (2020) ‘A Comparative Analysis Of The Factors Affecting Happiness Index’, 9(03).
2. Viet, C. (2021) ‘www.econstor.eu’, Does the COVID-19 Pandemic Cause People to Be Unhappy? Evidence from a Six-Country Survey. Available at: http://hdl.handle.net/10419/228738.
3. Kwak, S. K. and Kim, J. H. (2017) ‘Statistical data preparation: Management of missing values and outliers’, Korean Journal of Anesthesiology, 70(4), pp. 407–411. doi: https://doi.org/10.4097/kjae.2017.70.4.407.
4. Haiti | History, Geography, Map, Population, & Culture | Britannica (no date). Available at: https://www.britannica.com/place/Haiti (Accessed: 14 December 2021).
5. Linear Regression. Back to Basics. | by Sanchit Minocha | Data Science Group, IITR | Medium (no date). Available at: https://medium.com/data-science-group-iitr/linear-regression-back-to-basics-e4819829d78b (Accessed: 13 December 2021).
6. A Gentle Introduction To Neural Networks Series — Part 1 | by David Fumo | Towards Data Science (no date). Available at: https://towardsdatascience.com/a-gentle-introduction-to-neural-networks-series-part-1-2b90b87795bc (Accessed: 13 December 2021).
7. Perez-Vincent, S. M. et al. (2020) ‘COVID-19 Lockdowns and Domestic Violence: Evidence from Two Studies in Argentina’, inter- American Development Bank, pp. 1–48.
8. Waters, L. et al. (2021) ‘Positive psychology in a pandemic : buffering, bolstering , and building mental health’, The Journal of Positive Psychology, 00(00), pp. 1–21. doi: https://doi.org/10.1080/17439760.2021.1871945.
9. COVID-19 cases, recoveries, and deaths | Statista (no date). Available at: https://www-statista-com.manchester.idm.oclc.org/statistics/1087466/covid19-cases-recoveries-deaths-worldwide/ (Accessed: 12 December 2021).
10. Medical Association, A. (2020) ‘Psychological Distress and Loneliness Reported by US Adults in 2018 and April 2020’. doi: https://doi.org/10.1001/jama.2020.9740.
11. Fisher, J. R. et al. (2020) ‘Mental health of people in Australia in the first month of COVID-19 restrictions: a national survey’, MJA 213 (10) ▪ 16 November 2, 213(10). doi: https://doi.org/10.5694/mja2.50831.
12. Prime Minister confirms move to Plan B in England - GOV.UK (no date). Available at: https://www.gov.uk/government/news/prime-minister-confirms-move-to-plan-b-in-england (Accessed: 12 December 2021).
13. Helliwell, J. F., Layard, R., et al. (2021) ‘World Happiness Report 2021 | World Happiness Report’, World Happiness Report, p. 212.
14. Paper, W. and Paper, D. (2020) ‘www.econstor.eu’, Working Paper Happiness-lost: Did Governments make the right decisions to combat Covid-19?
15. Ju, Y. et al. (2020) ‘China’s mental health support in response to COVID-19 : progression, challenges and reflection’, pp. 1–9.
16. Shi, G. et al. (2021) ‘Factors influencing protective behavior in the post-COVID-19 period in China : a cross- sectional study’, 2, pp. 1–12.
17. Design, M. and Europe, W. (2007) ‘Gallup World Poll Research Design’, pp. 2006–2007.
18. Helliwell, J. F., Huang, H., et al. (2021) ‘Statistical of World Happiness Report 2021’, pp. 1–51.