HAPPINESS SCORE IDENTIFICATION: A REGRESSION APPROACH

Yichen Ma[1], Andrew Liu[2], Xukai Hu[3], Yuchen Shao[4]

[1]The Barstow School, Kansas City 64114, US
[2]Hpokinton High School, Hopkinton 01748, US
[3]Department of ECE, Virginia Tech, Blacksburg 24060, US
[4]Changzhou No.2 Middle School, Changzhou 213003, China

ABSTRACT: Happiness plays an important role in human emotion and one's growth. In this paper, we use the data from the *World Happiness Report*, *Countries of the World*, and *Countries Dataset 2020* to discern the relationship the happiness score has with the economy, family, health, freedom, trust, perception of corruption, generosity, and residual. In our research, we used the regression approach to find the most important factors that affect the happiness score in the past five years. Since we observed a positive and moderate relationship between the residual and happiness score, we then looked for other factors that contribute to the residual, the unexplained factor. Finally, we verified the main factors to the happiness score.

1 INTRODUCTION

Happiness has been discussed and pursued since the birth of mankind. Ancient Greek philosophers had many unique insights on happiness; for instance, Aristotle once said: "Happiness is the ultimate goal of all our actions, we do all things are in fact the means [1]." Today, when science and technology are highly developed, materialism no longer bothers and occupies people's minds. More and more people have begun the pursuit and discussion of happiness [2]. However, discussions that lack scientific and factual basis are meaningless, and differences in personal circumstances discern everyone's standards and perceptions of happiness. The United Nations (UN) has published many research reports on the happiness index [6]. Though, we found that these reports are simply the addition of the scores of all aspects. The reports did not find the factor that has the greatest impact on the happiness index. We decided to make up for the lack of conclusions and research in this regard.

We used the regression approach to explore the major influencing factors. In our opinion, the regression itself has a series of advantages such as simple operation and intuitive results. In addition, different countries have their own unique national conditions, and the major events that occur in the country each year are not the same. In our research process, we first did a rough and comprehensive analysis of all countries through the data analysis function of Excel, and pre-categorized the research results according to different regions, getting a preliminary conclusion: for most countries, economic it is a well-deserved factor that has the greatest impact on the happiness index. However, as we said before, national conditions and major events will have a major impact on the happiness index, this view is most obvious in Benin, Venezuela, and Algeria. They may encounter political turbulence or security threats, which not only caused various levels of happiness scores and total scores to decline, at the same time, the economy is no longer the most important factor affecting the happiness index.

According to the preliminary study mentioned above, our research provides a good summary of the factors that have the greatest impact on the happiness index. with only abstract data and the inability to conduct field surveys, we only have one factor of why the economy is the most influential.

The paper is organized as follows. The regression approach and the process of how to reach the final conclusion are described in Sections 2 and 3. Section 4 gives experimental results and conclusions obtained by combining the results of regression and machine learning, and Section 5 concludes our observations.

2 PREVIOUS WORK

In [3], based on the field survey of 578 elderly people in China and Japan, the authors used factor analysis, correlation analysis, t-test, and other statistical methods to find out the factors that affect the elderly's happiness and the differences between the two countries. Therefore, they found that the factors affecting the elderly's subjective well-being are not single, but multiple, including health factors, family factors, material life factors, different social environment, and cultural background will create

Email of all the authors: ajaxyichen@gmail.com, ajax.ma@barstowschool.org, andrewliu026@gmail.com, Guzh1MuXing@gmail.com, SYC9108@163.com

different happiness models; the happiness of the elderly needs not only material security but also spiritual security

In [4], through the coding and content analysis of written materials, this study explores the real concept of family happiness of ordinary people by combining qualitative and quantitative analysis. The coded views on family happiness were classified into 16 categories, of which harmony and solidarity were the most emphasized. Health and safety and income worry were the second and third respectively. Only 3% of respondents regarded money as the only element of happiness. The results of cluster analysis classified the expression of family happiness into three categories: "feeling good" (feeling orientation), "harmonious coexistence" (relationship orientation), and "economic security" (condition orientation). Among them, identity relationship orientation is the most common, and economic condition is regarded as about 1 / 3 of the elements of family happiness. Hedonism does not occupy the mainstream in the real world. Both qualitative and quantitative data do not support that ordinary people's view of family happiness is dominated by materialism and economic interests, but those with the poor economic background are more likely to identify with the economic security factors of happiness, and their family life is less happy.

In [5], the authors discussed how happiness differs from life satisfaction and proved that happiness can be measured by surveys. They showed that adaptation features well-being, and many common but important life events impact self-reported happiness profoundly. Yet adaptation to some events, such as long-term unemployment, is neither perfect nor immediate.

In contrast to these previous researches, our regression approach is simple and intuitive. We adopted a more representative sample of data from all over the world rather than selected countries that previous studies did, such as Japan and China. Also, when it comes to the happiness score, we did not restrain ourselves into the factors our source provided, but looked for other sources for other potential factors which elevates our credibility.

3 APPROACH

We used the dataset *World Happiness Report,* which contains the happiness scores, rankings, and potential factors across the past five years, from 2015 to 2019 [6]. The dataset adopts data from the Gallup World Poll. The poll asked living evaluation questions, known as the Cantril Ladder which asks respondents to rate their own lives of the time in a range of 0 to 10, and tallied the scores. The dataset contains the following attributes: countries, ranks, happiness score, economy (GDP per capita), health, family, trust, freedom, generosity, and dystopia residual. The adoption of dystopia aims to set the baseline of the lowest scores which avoids the occurrence of negative value, and the residuals are characterized as unexplained factors.

The regression approach features an interrelation of several independent variables, commonly known as x, and one dependent variable, commonly known as y, and offers an insight to see into the future trend. The performance of the regression lines of all the factors against the happiness score allows us to discover which factor contributes the most to the happiness score. The strength of correlation (a good, moderate, or bad correlation), the slope, and the R-square furthered our understanding of the factors. The question then shifted to the residual; a positive correlation with the residual and happiness score has been observed which means the residual could be further analyzed.

To have a better understanding of the residuals, we decided to look for other datasets for deeper analysis. Thus, *Countries of the World* grasped our attention immediately [7]. This dataset accommodates a variety of fields, such as population density, birth rate, literacy, industry, and so forth. However, when we performed both the linear regression and multiple regression (an approach that determines if several factors contribute to the independent variable as a whole) with the factors in the *Countries of the World* and the residuals in the *World Happiness Report*, bad correlations, very low R-squares, emerged on every regression graph we made. We then used Excel to filter the countries into different regions, trying to see if the bad correlation would still persist in a smaller sample that shared an identified characteristic. Again, the bad correlation thwarted our progression, while we discovered that the data in the *Countries of the World* were compiled by the US government in the period from 1970 to 2017. The content page of the dataset did not provide any additional explanation of the data nor the period. Therefore, we assume that the data took the average during this period which does not accommodate a specific year.

Another dataset *Countries Dataset 2020* saved us from the dilemma [8]. The dataset contained costs of living index, safety index, and healthcare index which are excluded in the *World Happiness Report*, albeit the data is gleaned in 2020. To minimize the error, we chose the happiness data in 2019 to perform regressions with; although the happiness data in 2019 and the *Countries Dataset 2020* only gaps half a year, the error still exists but has been mitigated. The regressions revealed a decent correlation with the happiness scores and contributed to the additional factors to the happiness score.

After discovering several essential factors to the happiness score, whether the countries that had a pleasant happiness score last year would remain happy this year intrigued us. Since the residuals and the happiness score have a moderate positive correlation, we realized that we could perform regressions of the happiness score from last year against the residual this year. The results correspond to our expectations. Following the findings that a relationship between the happiness score of last year and the residual of this year exists, we wanted to predict the 2020 residual values. We fitted each year's residual-happiness score data into a linear regression model and used this model to predict the 2020 residual value.

4 RESULTS

TABLE 1 Yearly R^2 Values for Factors vs. Happiness Score

	2015	**2016**	**2017**	**2018**	**2019**	**Average**
Economy	0.061	0.625	0.675	0.652	0.627	0.638
Health	0.524	0.586	0.607	0.594	0.607	0.584
Family	0.548	0.546	0.574	0.571	0.602	0.568

We used linear regression graphs to demonstrate the relationship between a country's happiness score and its other factors: economy, family, health, freedom, trust, and generosity. According to Fig. 1, since all the R^2 values for the given factors are all greater than 0.5, they were all very important to a country's happiness. Economy occupies the most significant factor to the happiness score with the highest R^2 value of 0.64. There was another factor whose relationship with the happiness score was quite strong, and that was the residual.

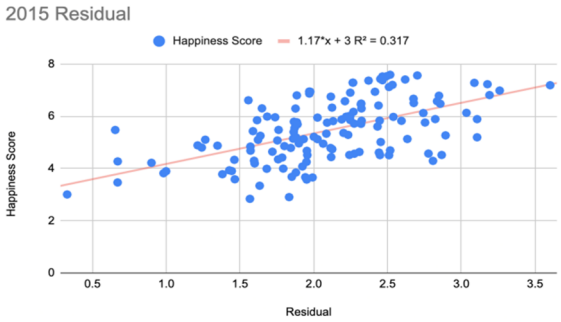

Fig. 1. The correlation between Residual and Happiness Score in 2015.

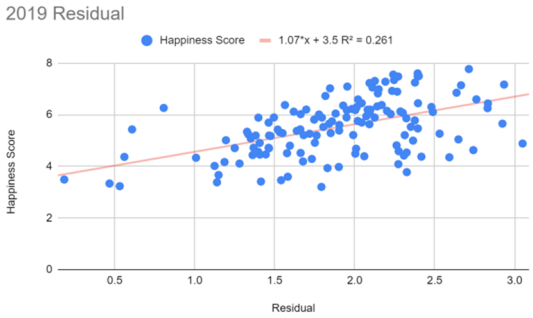

Fig. 2A. The correlation between Residual and Happiness Score in 2019.

Since the residual also seemed to correspond with the happiness score, we chose to look into what the "unexplained factors" could be and how they could help us predict this year's average happiness score, which is where we got these next regression plots.

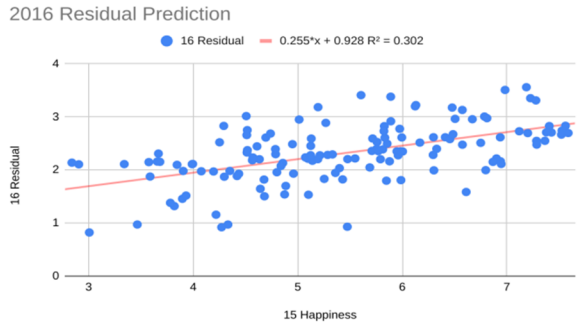

Fig. 3A. The correlation between Happiness Score in 2015 and Residual in 2016.

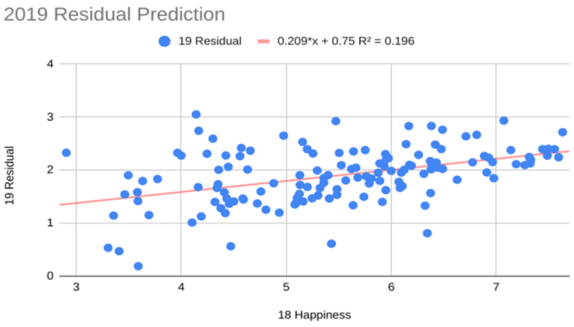

Fig. 3B. The correlation between Happiness Score in 2018 and Residual in 2019.

Finally, once we assured the fact that the residual maintained a strong relationship with the happiness score, we performed more regression on factors like population density, crime rate, etc. to find the greatest R^2 value, which would give us our possible candidates for the residual. As seen in Figure 4A-5B, two candidates ended up being the cost of living (plus rent) and the quality of the country's healthcare.

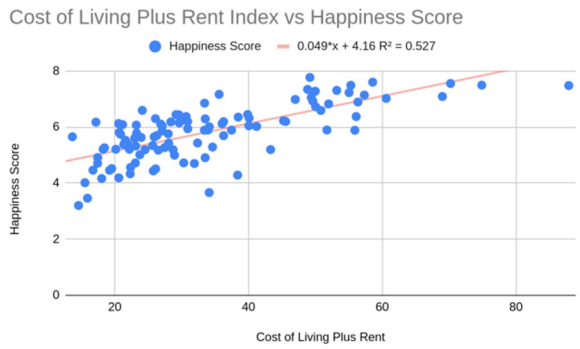

Fig. 4A. The correlation between The Cost of Living Plus Rent and Happiness Score.

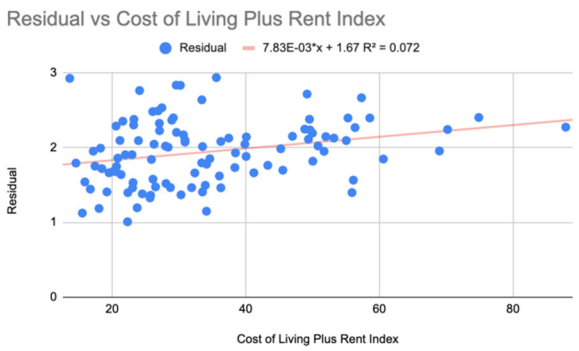

Fig. 4B. The correlation between The Cost of Living Plus Rent and Residual.

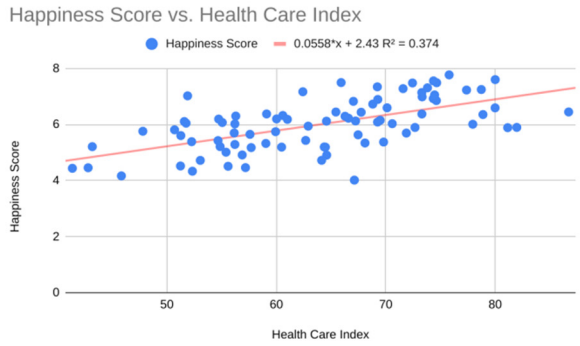

Fig. 5A. The correlation between Health Care Index and Happiness Score.

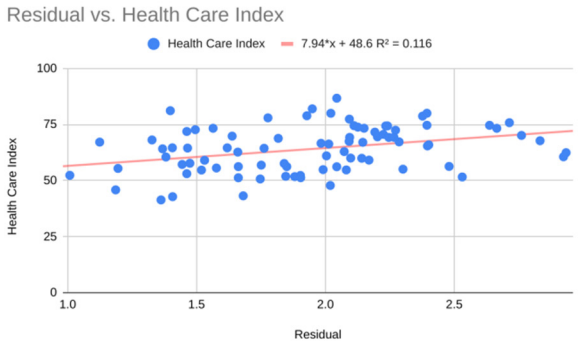

Fig. 5B. The correlation between Health Care Index and Residual.

5 CONCLUSION

In this paper, we presented the regression charts of economy, family, health, freedom, perception of corruption, and generosity with happiness from 2015 and 2019. Through comparisons between 2015 and 2019, we conclude that economy, family, and health are the most important factors that affect the happiness score. Especially, health's impact on happiness in recent years becomes much more important than in 2015. We utilized last year's happiness value to predict the next year's residuals value and found that a positive relation exists. In other words, people stay happy if they were happy last year. We called this phenomenon - happiness momentum - and this momentum plays an important role in each year's residuals.

The R-square value of the economy stayed around 0.62 in the past five years. The R-square value of the family with happiness increased by 0.04 from 0.562 in 2015 to 0.602 in 2019. The R-square value of the health with happiness increased by 0.064 from 0.543 in 2015 to 0.607 in 2019. We consider the economy as the most important factor that affects a person's happiness, but people are starting to pay more and more attention to family and health in recent years; in the future, health and family may become the most important factors in happiness.

In the future, we planned to make a new model for calculating the happiness score which would replace the model our source used for constructing the dataset. The model our source used only concluded 5 factors and a residual, which made the residual, the unexplained factor, becomes one of the major factors to the happiness score. Nevertheless, our research discovered 2 new factors, the cost of living plus rent index and healthcare. If we will be able to make the new model, the residual will be highly deducted which can cause the happiness score to be more realistic and accurate. Moreover, we planned to employ a neural network for further our research.

REFERENCES

1. Epicurus. "To live is to pursue happiness and happiness." 341 B.C- 270B.C.
2. Pan, Xiaodong. "From pursuing material to pursuing spirit." Printing Field 03-2018.
3. Lei, Xiuya. "A study on the subjective well-being of the elderly." Social Science Research 06-2004.
4. Erinys. "Family happiness: is money more and more important." Social Science research 01-2011.
5. Ortiz-Ospina, Esteban, and Max Roser. "Happiness and Life Satisfaction." *Our World in Data*, 14 May 2013.
6. Network, Sustainable Development Solutions. "World Happiness Report." *Kaggle*, 27 Nov. 2019.
7. Lasso, Fernando. "Countries of the World." Kaggle, 26 Apr. 2018.
8. Yadav, Varun. "Countries Dataset 2020." Kaggle, 21 Mar. 2020