THE CANADIAN SCIENCE FAIR JOURNAL

ARTICLE

Fountain of Youth

By: Angela Cao | Edited by: Philippe Nadeau | Layout by: Ahmed Nadeem

Age: 13 | Burnaby, BC

Canada-Wide Science Fair 2022 Junior Silver Medal, $2000 Western University Science Entrance Scholarship, GVRSF 2022 Junior Gold Medal, GVRSF Junior UBC Statistics Award, GVRSF SFU Faculty of Science Award

From the 5th century BC to modern times, society has been obsessed with immortality and the unknown premonition of death. The primary objective of this project was to analyze the relationships between contextual influences (independent variables) and life expectancy using the simple linear regression model. Secondary objectives include developing a RShiny application to predict a country's average life expectancy through user-input values and visualize the independent variables changing over the past years through choropleth maps. The selection of independent variables: happiness score, Per Capita GDP (PCGDP), and average years of education are supported by conclusions from Kabir (2008), which states developing countries should increase adult literacy and nourishment to improve lifespan. Further, studies have shown that happier people have greater lifespans (Argyle, 1997; Deeg and van Zonneveld, 1989; Howell et al., 2007).

METHODS AND MATERIALS

2.1 Data Sources

This study is on a national level, and the estimates should not be considered equivalent to ones done on an individual level. Primarily, four datasets were used: two datasets (life xpectancy and happiness) from an online community for data science (Kaggle, 2019) and two datasets (PCGDP and education) from The World Bank (World Bank, 2021). Although happiness is an abstract concept, in this study, it is determined by the World Happiness Report (World Happiness Report, 2019), which computes a measure of the happiness score of each country from the Gallup World Polls. Average years of education refers to the average years of education of the adult being in that country; the data is from the Barro-Lee Educational Attainment Data.

2.2 Data Cleaning

The datasets differed greatly from one another with additional rows or columns of data that did not occur throughout all four and thus were removed. Originally, the four datasets had a total of 8,427 rows and 134 columns. After removal, the final dataset totalled 144 rows and 15 columns. Moreover, the datasets were individually reorganized to remove excess headers.

Because life expectancy, happiness scores, average years of education, and PCGDP are not constant across different years but rather change, only data from 2017 was used to control the study. (2017 was the most recent year that had available data across all four variables.) In addition, the data for the "continent" columns used in Fig. 2 was manually input.

2.3 Pearson Correlation coefficient

The following Pearson correlation coefficient r is used to measure the correlation and linear relationship between two variables:

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

and x_i and y_i are the means of the values of the x-variables and y-variable, respectively. The Pearson correlation coefficient has a range of -1 to +1. A large absolute value of *r* (typically greater than 0.7) indicates a strong correlation.

2.4 Modeling

The linear regression model in this study is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where Y_i is the dependent variable, X_i is the independent variable, β_0 is the y-intercept, β_1 (slope) is the change in variable Y_i when the variable X_i changes by one unit, and ϵi is the random error following a normal distribution (mean 0, variance σ^2).

RESULTS

3.1 Analysis

The last row of Table 1 shows the Pearson correlation coefficient (see Section 2.3) between life expectancy and the three independent variables. A logarithmic transformation of PCGDP was done because its correlation with life expectancy was stronger than using the original scale. Table 1 shows that the correlation between log PCGDP and life expectancy (visualized in Fig 1)is the strongest (i.e., 0.84) of the three independent variables studied.

A linear regression was performed with the model defined in section 2.4 on data from happiness, log PCGDP, and education, respectively. As seen in Table 1, all three estimated slopes are positive, meaning that if the variable increases by one unit, the life expectancy increases by its slope. To visualize the relationship between the independent variable and life expectancy, slope

Table 1: The estimates for the linear regression model from the collected data including y-intercept, slope, 95% confidence interval, p-value, and Pearson correlation coefficient.

Factor	Happiness	Education	Log PCGDP
Y-intercept	43.60	56.12	19.17
Slope	5.40	1.89	13.05
95% Confidence Interval	[4.71, 6.09]	[1.66, 2.11]	[11.75, 14.34]
p-value	< 0.001	< 0.001	< 0.001
Pearson Correlation Coefficient	0.79	0.77	0.84

and y-intercept can be used to write an equation for each variable in slope-intercept form:

$$\text{Life Expectancy} = 43.6 + 5.4 \times \text{Happiness}$$
$$\text{Life Expectancy} = 56.1 + 1.9 \times \text{Education}$$
$$\text{Life Expectancy} = 19.2 + 13.0 \times \log(\text{PCGDP})$$

Even in modern times, Fig 2 visualizes the drastic differences between continents. African countries that may be less developed, consistently appear in the bottom left corners which further agrees with Kabir (2008). In contrast, European countries appear frequently in the top right corners.

Furthermore, to find whether these results are statistically significant, statistical tests were conducted to test the null hypothesis, which can be stated as there is no linear relationship (slope=0) between the two variables. In contrast, an alternate hypothesis can be stated as there is a linear relationship. If the p-value is less than 0.05, the results are statistically significant and give evidence as to why the null hypothesis is wrong as there is an unlikely probability (<0.05) of obtaining results when the null hypothesis is correct. Table 1 shows that the three tests are statistically significant with a p-value of $<2.2e^{-16}$.

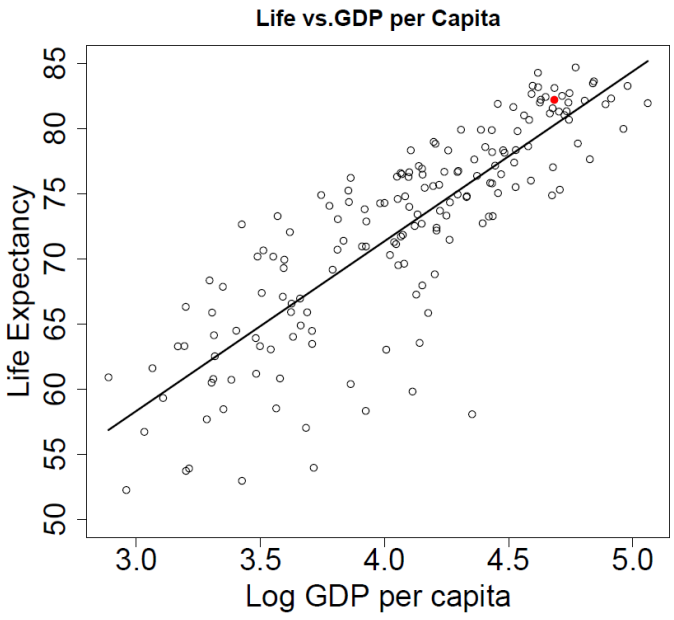


Figure 1: The relationship between log PCGDP and life expectancy. The red dot is Canada.

3.2 RShiny application

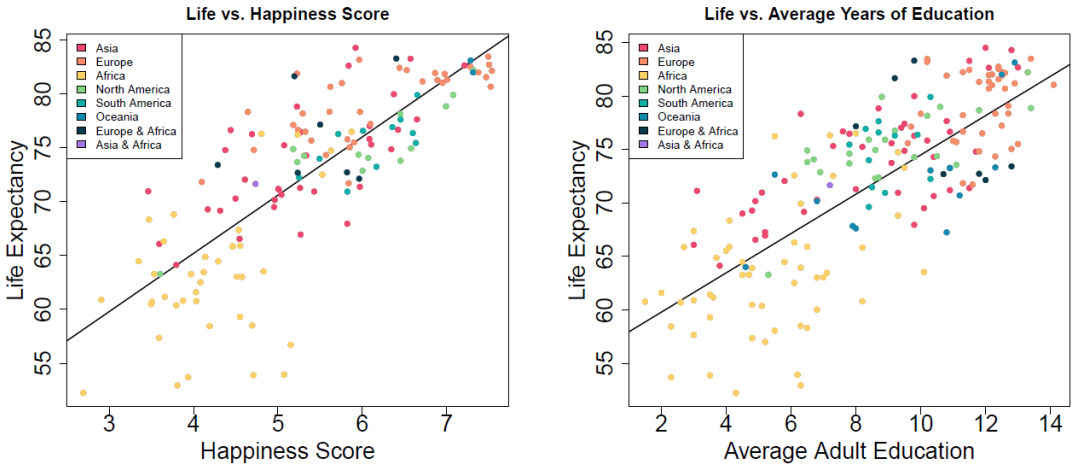
Using R (R Core Team, 2021), an RShiny application was developed to predict a country's average life expectancy based on user-input values, as shown in Fig. 3. The app fits multiple linear regression using all of the 3 independent variables together to predict life expectancy. Further, the application includes 38 interactive choropleth maps to visualize the global development of log PCGDP and average years of education over the past years (2002-2017). However, because the happiness dataset only included data for 2015-2017, there are only 3 maps depicting happiness scores, as shown in Fig. 4.

DISCUSSION

It is important to remember that correlation does not mean causation. Although all the variables are correlated with life expectancy, it does not necessarily mean the variables actively cause life expectancy but are rather only shown to be moving together. Further, it is assumed in this study that average years of education, happiness score, and PCGDP are independent variables. For future studies, interaction between these variables can be considered and tested through model selection methods.

Moreover, it is recommended to examine how life expectancy has developed over the past years. In this study, lifespan is predicted by 2017 data; however, it is probable that the datahas changed and certain factors could be more or less correlated. The RShiny application could be less applicable as time passes.

In addition, it could be beneficial to examine life expectancy on an individual level. This could result in more customization and expansion of the application. Other individual influences on lifespan could also be added to the study, including but not limited to, culture and individual financial backgrounds. Further, it could be useful to examine life expectancy and other factors on a continental level, including the variance of the data per continent.



(a) The relationship between happiness score and life expectancy for 145 countries when sorted by continent.

(b) The relationship between average years of education and life expectancy for 145 countries when sorted by continent.

Figure 2: There is a clear difference between the continents when comparing happiness scores and years of education.

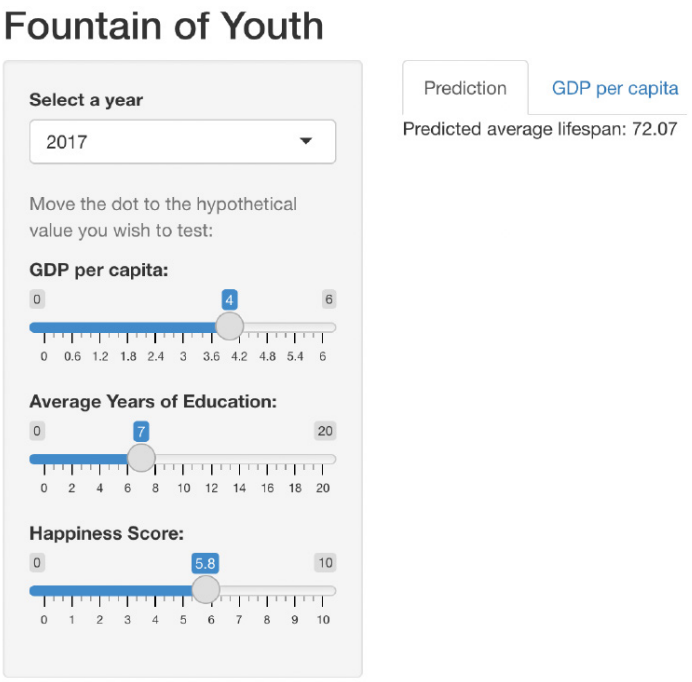


Figure 3: Demonstration of the RShiny application to predict a country's average life expectancy based on values the user inputs.

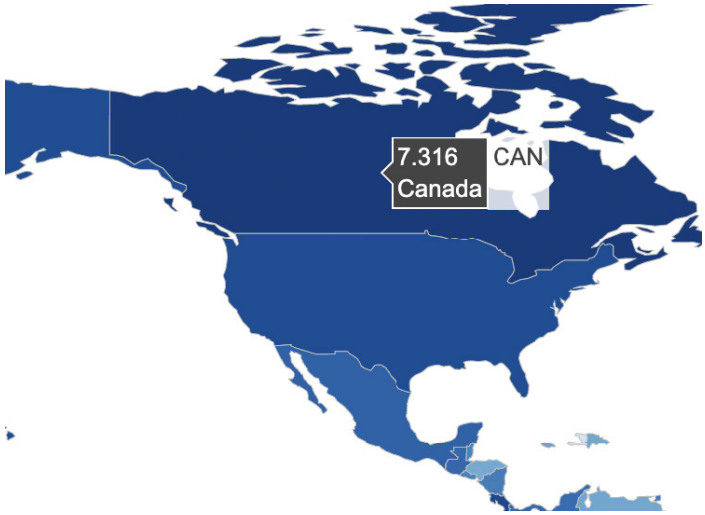


Figure 4: Zoomed in image of Canada from the 2015 happiness score interactive choropleth map.

CONCLUSION

The analysis of life expectancy is essential to further advance the understanding of human development, maximize lifespan, and create a productive society. The RShiny application takes a novel approach to life expectancy prediction as it takes into account present variables whereas predictions solely based on increases over the past years can be less precise. Because life expectancy plays an influential part in socioeconomic-demographic decisions, countries
need to know their expected life expectancy to calculate healthcare costs and program funding. Further, this study provides a strong foundation for future research on life expectancy and lifespan development.

REFERENCES

Argyle, M. (1997). Is happiness a cause of health? Psychology & Health, 12 (6), 769–781. https://doi.org/10.1080/08870449708406738

Deeg, D. J., & van Zonneveld, R. J. (1989). Does happiness lengthen life? The prediction of longevity in the elderly. How harmful is happiness, 29–43.

Howell, R. T., Kern, M. L., & Lyubomirsky, S. (2007). Health benefits: Meta-analytically determining the impact of well-being on objective health outcomes. Health Psychology Review, 1 (1), 83–136. https://doi.org/10.1080/17437190701492486

Kaggle. (2019). Kaggle: Your home for data science. https://www.kaggle.com/

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation

for Statistical Computing. Vienna, Austria. https://www.R-project.org/ World Bank. (2021). World Bank Group - International Development, Poverty, Sustainability.

https://www.worldbank.org/en/home

World Happiness Report. (2019). Home. https://worldhappiness.report/

ABOUT THE AUTHOR - ANGELA CAO

Angela Cao is currently attending the University Transition Program in Vancouver, British Columbia. She developed an interest for statistics after discovering its possible applications to a wide array of real-world problems and plans to pursue it through future projects. Further, she hopes to pursue medicine during post-secondary education and aims to contribute discoveries in the field of health sciences and deepen our understanding of the human body. After taking on a huge scientific project for the first time for the 2022 CWSF, she is excited to attend future STEM fairs and present new findings and innovations.