April 27th, 2023

CS555 – Term Project

Zuowen Tang

Data Science Research

On Humanity and Happiness

Based on World Happiness Report from 2017-2019

# Introduction

Every individual might have different goals, ambitions, and purposes in their life, yet one thing is what we all want: happiness. No matter who you are, an intelligent young scientist who wishes to work at NASA, a beautiful dancer who wants to be on a big stage, or an old gentleman who only wishes to spend more time with his family, happiness is always the thing that attracts us. Today we live in a society where everything seems able to be quantified, knowledge, feelings, personal information, and maybe happiness as well. Starting in 2011, the World Happiness Report began to gather nationwide data based on scientific facts and *subjective ratings from respondents based on their own life[[1]](#footnote-1),* with various life factors, to see which country is the happiest. On the other hand, this data set could also answer these questions: what determines people's happiness? Under what circumstances can we live the best life? Also, based on the data set we already have, can we predict happiness in the future?

# Data Set Description

This data science research is based on the dataset of the World Happiness Report from 2017-2019, which can be accessed on the public data set website *Kaggle* and *WHR's webpage*. There are six main variables that are considered important to people's life. *They include GDP per capita, social support, healthy life expectancy, freedom, generosity, and corruption[[2]](#footnote-2).*

Graphical user interface

Description automatically generated with medium confidence

Figure 1 Data set variables

This study will mainly consider variables such as **GDP per Capita**, **Social Support**, and **Healthy Life Expectancy** as the dominant factors since they strongly correlate with the overall score. **GDP per Capita** is *an economic metric that breaks down a country's economic output per person[[3]](#footnote-3).* **Social Support**, also known as family support in the past data set report, is more personal and subjective. It means how much emotional support a person can receive from their family/friends. This variable is rated by respondents based on their experience. **Healthy Life Expectancy** means *the average number of years that a person can expect to live in full health[[4]](#footnote-4)*. The other three variables, **Freedom** *(whether a person can make a decision freely),* **Generosity** *(how much/often people will donate to charity),* and **Corruption** *(the political corruption or the abuse of power based on people's experience),* will only be mentioned in very few specific tests.

# Methods Introduction

The following statistical methods will be used in this report to help the writer draw connections between factors and form a conclusion: two-sample mean tests, correlation tests, simple and multiple linear regression, and ANOVA tests. Diagrams, such as histograms, box plots, scatter, and three-dimensional plots, will also be used in pursuance of virtualizing the dataset. Furthermore, machine learning algorithms, for instance, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Naive Bayes Classifier, will be used for the predictive analysis.

# Results Report

**1. Overall Graphical Summary**

A screenshot of a computer

Description automatically generated with medium confidence

Figure 2 Basic summary to show min/max/mean of each variable

Chart, box and whisker chart

Description automatically generated

The above diagram is a boxplot to virtualize each variable's min, max, median, IQR, upper, and lower quartiles after removing all the outliers. We can see that GDP, Social Support, and Health are three variables with a bigger range than others.

Chart

Description automatically generated

The figure on the right side is a histogram to virtualize variables in a different way. We can see that everything other than the Overall Score and Year is slightly skewed on either the left or right side, which means the resources of those variables are often easy to access in some countries but not others, and vice versa.

Text

Description automatically generated with low confidenceChart, histogram

Description automatically generatedFrom the diagram on the left side, we can see that the Score is always close to the normal distribution no matter in which year. We can also test the overall score in the Shapiro test to make sure about this part. As you see, the p-value of the Shapiro test is smaller than 0.05, which means the score is normally distributed.

**2. Data Testing**

Background pattern

Description automatically generated Now we have some understanding about the variables; we want to do some tests to find out which of those variables are the ones that determine the overall score the most, i.e., what can make human beings more happy.

Text, letter

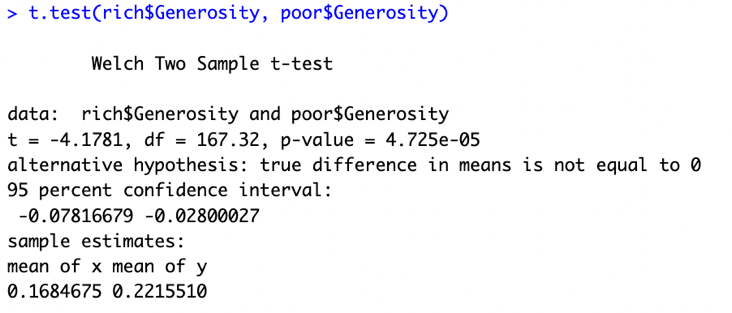
Description automatically generatedIn the correlation diagram and value list on the right side, we can see that GDP, Social Support, and Health have a correlation value of 0.75, 0.70, and 0.71 with the score. This means they are three main variables strongly related to the score. And with the check of linear regression and diagrams (see next page,) we can confidently say that all three of the variables have positive correlations with the overall score, which means the higher they are, the higher the overall happiness score will be. We will test some hypotheses based on this assumption.

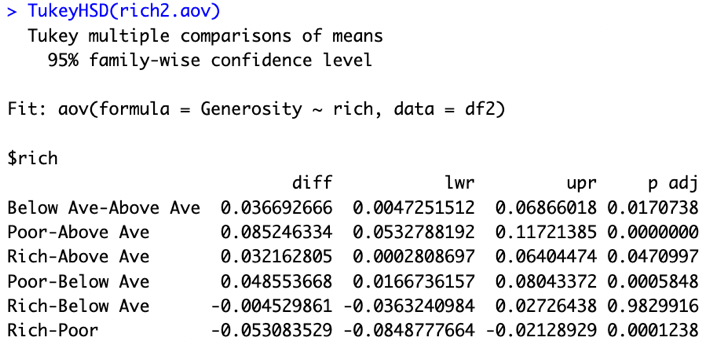
**Table

Description automatically generatedHypothesis**: from the above correlation diagram, we can see that there seems to exist a negative relationship between GDP and Generosity. Does this mean that people would donate less or less often when they are richer?

In order to test this, we first separate data into different groups: Rich, Above Average, Below Average, and Poor (based on the quartile.)

Figure 3 From the above linear regression results, the p-value is smaller than 0.05. This means that there is a linear association relationship. These three variables have a strong correlation with the score.

Based on the two-sample mean test between group rich and poor in the picture below, we can see that the p-value is smaller than 0.05, which means that it is statistically significant. The two groups do not share the same mean, and the mean value in the poor group is higher than the mean value in the rich group.

We can also check the difference between groups by using the Tukey HSD algorithm. The result below shows that groups between Poor-Above Ave, Poor-Below Ave, and Rich-Poor are statistically significant (p-value < 0.05.) From this test, the result shows that between rich and poor countries, the poor ones would normally donate more money or more frequently than the rich ones. While between countries that are above average and poor countries, or between rich countries and countries that are above average, this result is reversed. [[5]](#footnote-5)

Chart, scatter chart

Description automatically generated**3. Prediction Test**

The purpose of this section is to help scientists or reporter to predict happiness in the future. In order to accomplish it, we will be using LDA, QDA, and Naïve Bayes algorithms. The data will first be separated into three groups: **Happy**(3rd Q), **Okay**(between 3rd and 1st), and **Unhappy**(1st Q), based on the quartiles of the overall happiness score. The prediction will only use the three main variables: GDP, Health, and Social Support. In the 3D scatter plot on the right, we can see that although there are mixing results, there is still a rather obvious separation between the observations.

The following pictures are the prediction results based on the 8:2 ratio and 5:5 ratio, using LDA, QDA, and Naïve Bayes.

← The result on the left is based on 8:2 ratio of training and testing. The accuracy of each algorithm is 93% (LDA), 89%(QDA), and 86%(NB).

→ The result on the right is based on 5:5 ratio of training and testing. The accuracy of each algorithm is 94% (LDA), 89%(QDA), and 88%(NB).

A picture containing text

Description automatically generatedText

Description automatically generated with medium confidence

Figure 4 5:5 Ratio

Figure 5 8:2 Ratio

# Conclusion

From the above research, we might conclude that although money cannot buy everything, it is strongly related to people’s happiness. After all, a person who suffers from hunger and lack of resources cannot be happy. However, it is not the only factor since Health and Social Support are also heavily related to happiness. That might explain why some rich people also have to deal with depression and other sadness if they ever suffer from loneliness or illness. Of course, this research is based on a database created by World Happiness Report, whose calculation methods and algorithms are not publicly published. This could mean that without knowing how exactly they gather the data, the conclusion might not be significant enough for everyone living on the earth. There might be other factors that are important to human happiness that were not included in this dataset. Overall, happiness is a rather blurry concept since not everyone shares the same ideology of what happiness is. Still, if there is some idea that most people are all agreed on, statistically, we have found the definition of happiness.

***Dataset:***[*https://www.kaggle.com/datasets/unsdsn/world-happiness*](https://www.kaggle.com/datasets/unsdsn/world-happiness)

***Reference:***

*About | The World Happiness Report*. (n.d.). Home | The World Happiness Report. Retrieved April 27, 2023, from <https://worldhappiness.report/about/>

*FAQ | The World Happiness Report*. (n.d.). Home | The World Happiness Report. Retrieved April 27, 2023, from <https://worldhappiness.report/faq/>

PhD, M. S. (2007, May 30). *Healthy Life Expectancy and How It’s Calculated*. Verywell Health; Verywell Health. <https://www.verywellhealth.com/understanding-healthy-life-expectancy-2223919>

The Investopedia Team. (2009, July 8). *GDP Per Capita Defined: Applications and Highest Per Country*. Investopedia; Investopedia. <https://www.investopedia.com/terms/p/per-capita-gdp.asp>

1. *FAQ | The World Happiness Report*. (n.d.). Home | The World Happiness Report. [↑](#footnote-ref-1)
2. *About | The World Happiness Report*. (n.d.). Home | The World Happiness Report. [↑](#footnote-ref-2)
3. The Investopedia Team. (2009, July 8). *GDP Per Capita Defined: Applications and Highest Per Country*. [↑](#footnote-ref-3)
4. PhD, M. S. (2007, May 30). *Healthy Life Expectancy and How It’s Calculated*. Verywell Health; Verywell Health. [↑](#footnote-ref-4)
5. Other hypothesis that used exactly same methods for testing is omit in this report due to the length. The results of those tests can be found in the R script. [↑](#footnote-ref-5)