Exploring the Association Between Financial Indicators and the Happiness Level: A Country-Level Study

Big Data Analytics

(BYGB-7990-V02)

May 10th, 2021

Anuja Apte, Owakhela Kankhwend, Muhammad Faiq Nasir, Yue Wang, Jiacheng Zhang

**Executive Summary**

There is a common belief that economic crises will lead to a decrease in subjective well-being (happiness level). The aim of this research is to study if the financial uprise or downfall of a country affects the happiness level. The study will also focus on exploring which indicators influence the changes in happiness levels the most and the least.

Our dataset is collected from a variety of sources namely Gallup World Poll, World Bank, and World Happiness Report.

We will be using a combination of Big data techniques along with regression analysis. The two types of analysis we will be using are predictive analysis and descriptive analysis. The Logistic regression being a part of the predictive analysis helps us understand which financial indicators have a major impact on the happiness level of a country. Furthermore, the descriptive analysis gives us comparisons of financial indicators with the happiness indicators.

Our goal is to find out if the output of predictive and descriptive analysis gives similar output for indicating which variables impact Happiness level the most.

This study is carried out for a period of 13 years from 2006 to 2018 which was a pre-covid19 era. For future studies, it would be interesting to compare the happiness levels from pre-covid19 years to the active-covid years. Considering that all the countries worldwide have experienced an economic downfall in the pandemic, it would be fascinating to see if that has affected the happiness of people.

**Business Problem**

There are many times people come across the phrase ‘pursuit of happiness,’ however, what brings people happiness and how it is quantified is an issue. For this project, we aim to look at what indicators impact the happiness of a country. Many times, we hear people say, ‘Money doesn’t buy happiness.’ For this project, we want to evaluate the effects of money, which for this project would be the financial indicators, on the happiness level of a country.

In addition to this, the public becomes more and more interested in how to rank happiness among different countries and get to know what facts will influence their happiness ranking among different countries. What is exactly their right mix of ingredients for happiness? High GDP per capita, social support in times of need, absence of corruption in government, healthy life expectancy, freedom to make life choices, generosity, or democratic quality. What are the most important variables and what are the least important ones?

**System Design**

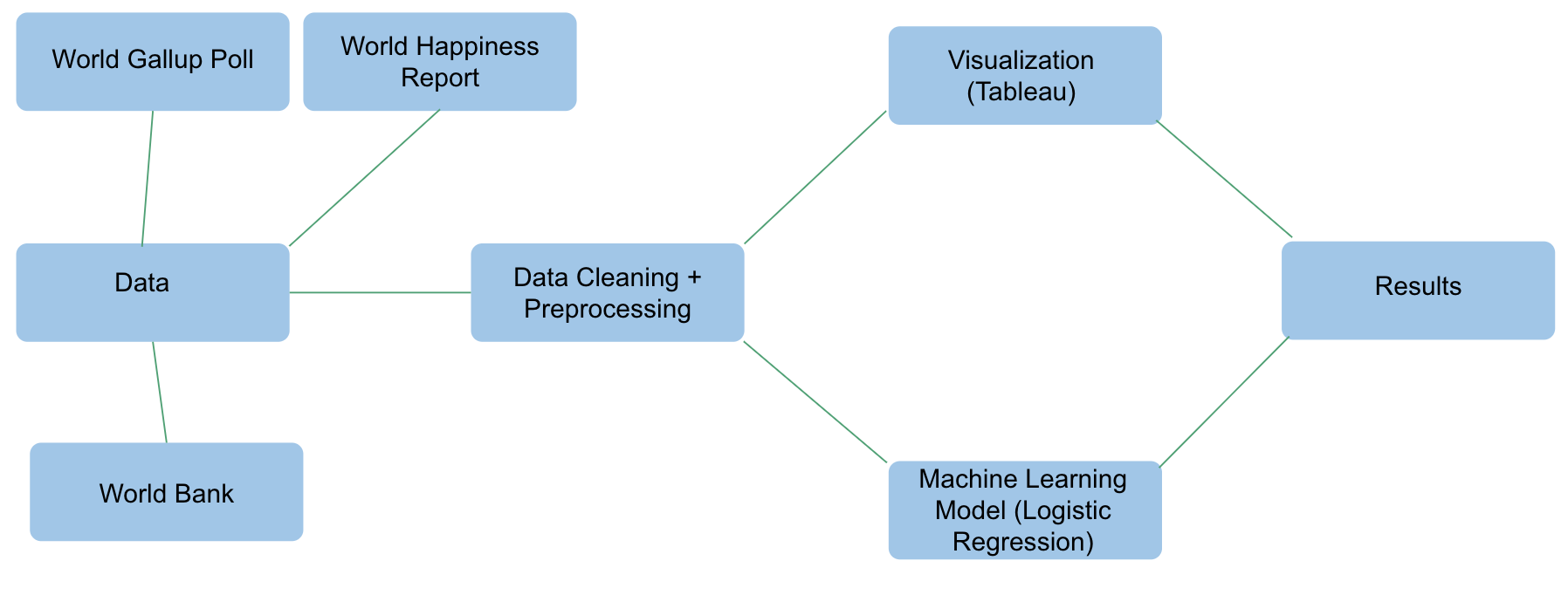
****

Figure 1: System Design

Figure 1 displays the system design for the project. The project was started by acquiring data from three different sources, the World Bank, the Gallup World Poll, and the World Happiness Report. Then the next step was Data Engineering, in which we cleaned the dataset in excel and imputed missing values. Next, the data was preprocessed in Google Cloud Platform using, jupyter notebook for python so it was ready to be used for descriptive and predictive analysis. We used Logistic Regression for our predictive machine learning model. For descriptive analysis Tableau was used to create descriptive figures. Lastly, both of these models are used to get the results of the analysis.

**Data Description**

The data was collected from three different sources, the World Bank, the World Happiness Report, and the Gallup World Poll. The data from the World Bank consists of different financial and other development indicators such as GDP per capita, GNI, and Total Reserves. The data from the Gallup World Poll and the World Happiness report consist of the happiness ladder, which essentially is the happiness score assigned to a country based on the survey response, where people rank from 0 to 10 where their happiness lies on the ladder. In addition to this, both the Gallup World Poll and the World Happiness Report were also used to find variables such as Social Support and Generosity that were based on poll responses. Once all the data was merged, there were around 75,653 rows that were cut down to 1559 rows due to missing values. In addition to this, the data set initially had 215 countries, however, due to missing values for the Happiness Ladder and other indicators the countries were brought down to 145 countries.



Table 1: Snapshot of the Final Dataset

Figure 2 shows the final dataset that contains 12 different variables. The independent variables include Log of GDP per capita, Social Support, Healthy life expectancy at birth, Freedom to make life choices, Generosity, Perceptions of corruption, GNI per capita, PPP (current international $), and Total reserves (includes gold, current US$). The control variables include Country Name and Year. And lastly, the dependent variable is the Happiness Ladder. For the purpose of analysis, the Happiness Ladder is used first in its actual form, which are 0 to 10 values recorded in the poll responses. Second, the score was converted into categorical variables where any score of 5 or above is ‘yes’ which means those countries are happy. However, any score below 5 will convert into ‘no’ which means those countries are not happy.

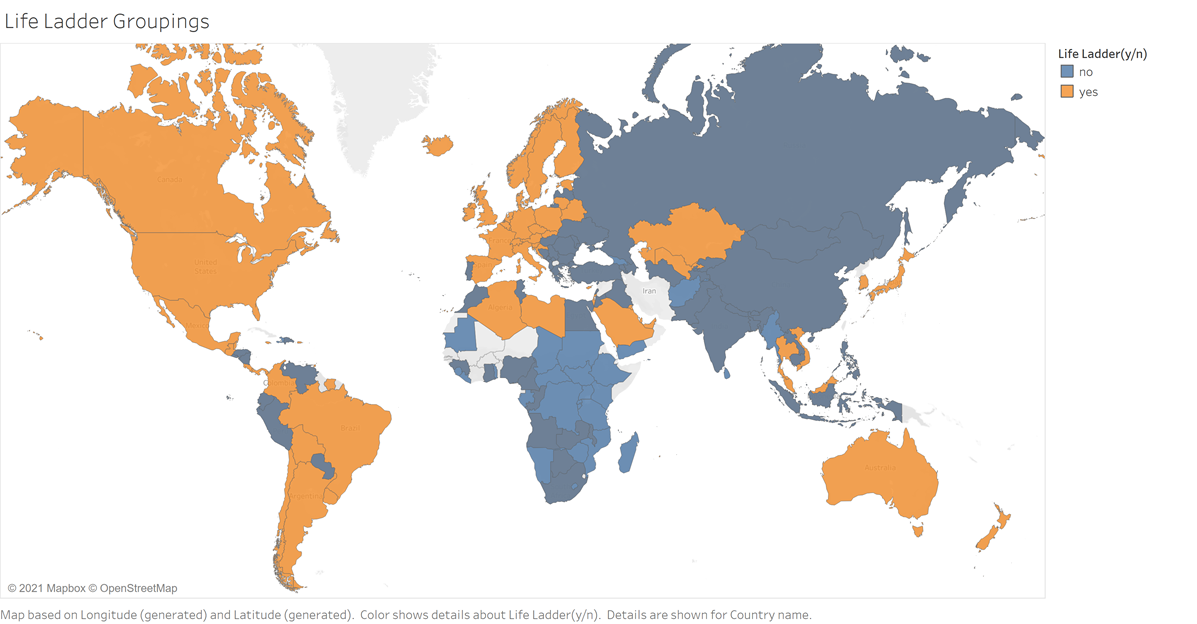
**Data Cleaning & Preprocessing**

As for the first step, our group collected raw data from the World Bank, the World Happiness Report, and the Gallup World Poll and merged all the data in excel. Once the data was merged, our next step was to clean the data and impute missing values. We filtered the raw data to remove unnecessary data that didn’t align with the rest of the data we had. Then we imputed missing values by using mode and median values to replace them. Our next step was to use google cloud computing and use jupyter notebook for preprocessing the data in python. Here we divided out variables into categorical and numerical variables. Then we standardized the numerical variables, while for the categorical variables we created redundant dummies. Now the data was ready for both predictive and descriptive analysis. For our logistic regression model, we partitioned the data into test and train data, 80 percent and 20 percent respectively.

**Descriptive and Predictive Analytics Visualizations**

We investigated the relationships between each of our independent variables and the Dependent variable, Life Ladder, in both its numerical form and its nominal form. The aim was to gain a thorough understanding of the association between each pairing, insight into the variables themselves, and insight into helpful recommendations for governments seeking to improve their life ladder score.

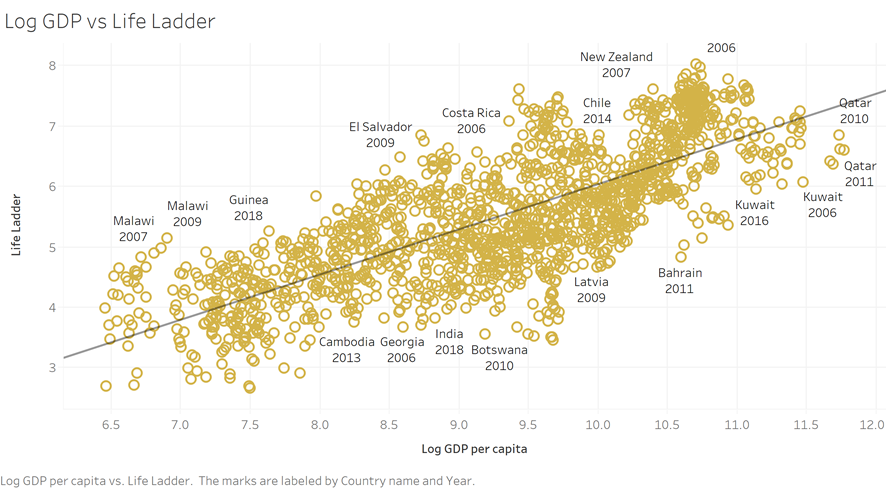
**Life Ladder**

****

**Figure 1: Map depicting the Life Ladder Groupings**

The map details the countries included in the study and the Life Ladder groups that they belong to. The “Yes” group are those with an average life ladder score above 5 over the course of the dataset timeline. The “No” group are those with an average life ladder score below 5 over the course of the dataset timeline. Notably, the “Yes” group appeared to be mostly made up of more developed countries while the “No” group appeared to be mostly made up of the lesser developed countries with notable exceptions of powers like Russia, China, India, and South Africa. As we hypothesized at the onset of the study, development is highly likely to be positively correlated with happiness scores.

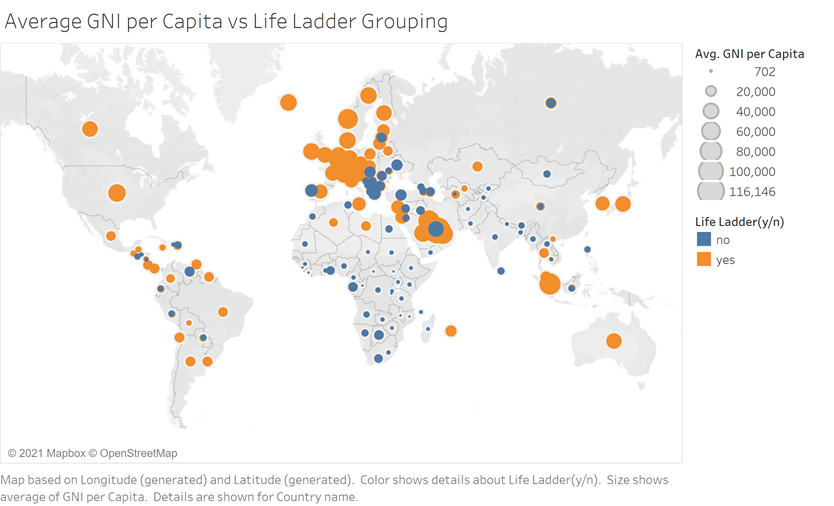
**Log GDP per Capita vs Life Ladder**

****

**Figure 2: Scatterplot depicting Log GDP per Capita vs Life Ladder Score**

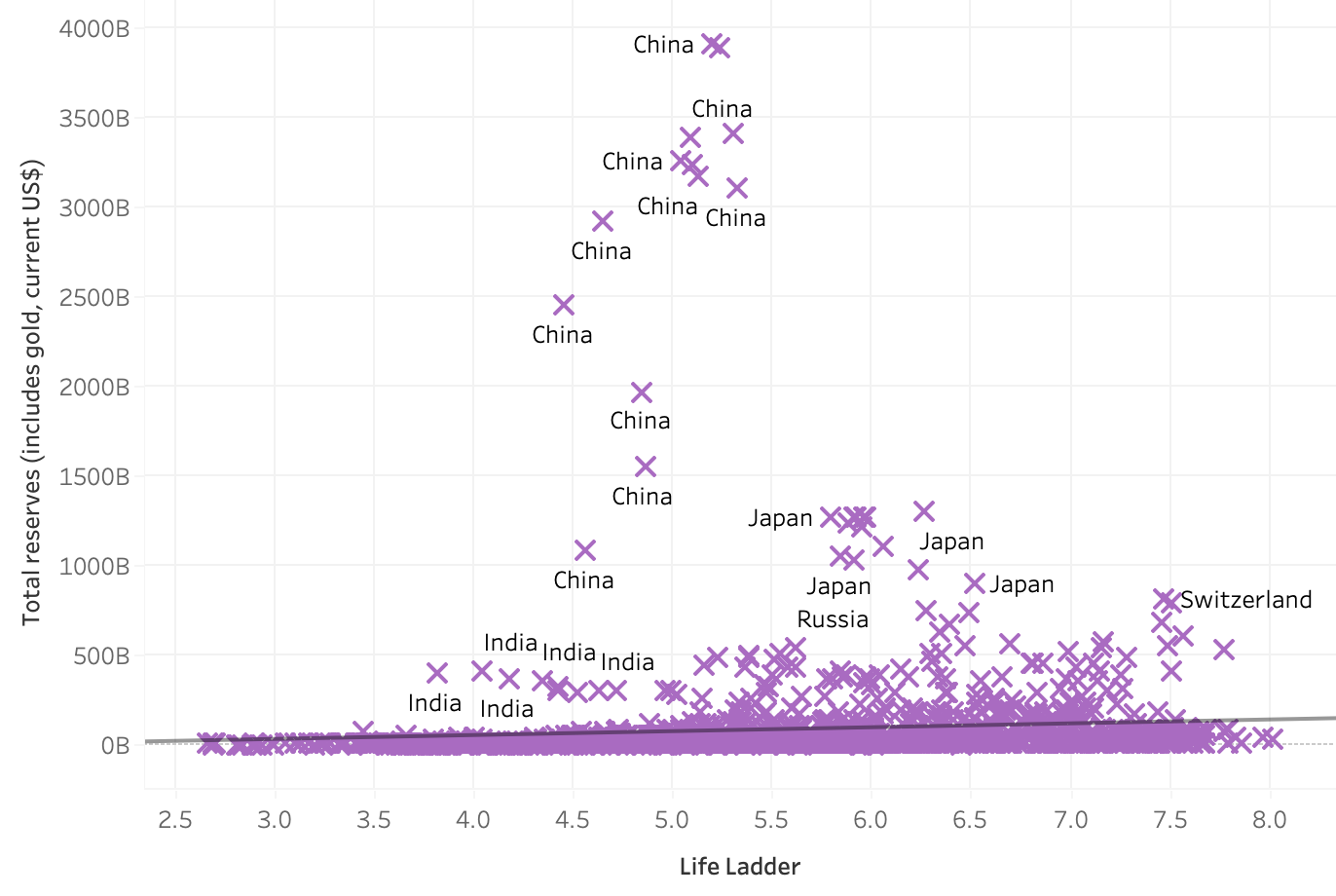
The scatter plot details the relationship between Log GDP per Capita and Life Ladder score with details of Country and Year for some of the observations. There was a strong positive relationship defined by the trendline formula of 0.748265\*Log GDP per capita - 1.451. Log GDP per capita has great value as a predictor given p-value < 0.0001. The regression model has a good value given an R-Squared measure of 0.600911. The insight gained is Log GDP per Capita is in fact a string descriptive factor on bettering the environment within a country to in turn foster a stronger happiness score. It is also a useful predictor in determining happiness scores.

**GNI per Capita vs Life Ladder**

**Figure 3: Map depicting Average GNI per Capita vs Life Ladder Grouping**

The Map shows the relationship between the Average GNI per Capita for each country over the course of the dataset years versus the Life Ladder Grouping. The map details a positive relationship with the “Yes” group more commonly having a larger Avg. GNI per Capita for all years in the dataset. The implication is that GNI per capita is a reliable associative variable for Life Ladder scores in the dataset.

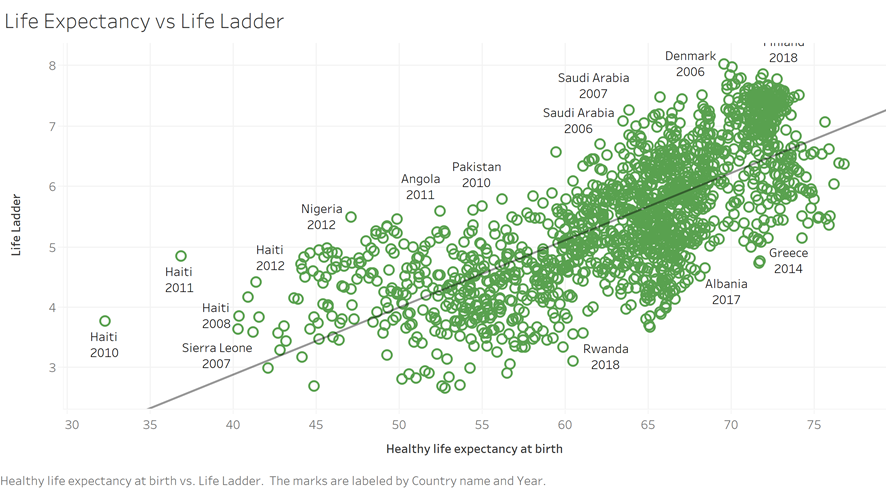
**Total Reserves vs Life Ladder**

**Figure 4: Scatter plot for Total reserves vs Life ladder**

This graph shows the relationship between the Life ladder and the Total reserves of a country. It is clear that the Total reserves of a country do not influence its happiness level. Contrary to our beliefs, it is evident that most of the countries with a higher total reserve result in a low life ladder score. Hence there doesn’t seem to be any strong correlation between the two variables.

Switzerland and Japan seem to be the only exceptions for this analysis.

**Healthy Life Expectancy at Birth vs Life Ladder**

**Figure 5: Scatter plot depicting Healthy Life Expectancy at Birth vs Life Ladder Score**

The scatter plot details the relationship between Healthy Life Expectancy at Birth and Life Ladder score with details of Country and Year for some of the observations. There is a strong positive relationship defined by the trendline formula of 0.11132\*Healthy Life Expectancy at Birth -1.57483. Healthy Life Expectancy at Birth has great value as a predictor given p-value < 0.0001. The regression model has a good value given an R-Squared measure of 0.537095. The insight gained is Healthy Life Expectancy at Birth is positively associated with a stronger happiness score. It is also a useful predictor in determining happiness scores.

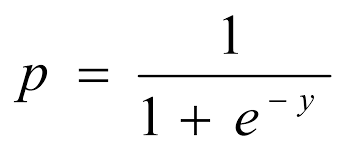
**Logistic Regression**

When we consider our business problem, we believe that using a predictive model to predict the potential financial indicators would be helpful for this project. Also, after finishing the data preprocessing and descriptive analysis, we were able to present a further machine learning approach and algorithm that gives us more features related to the topic we were looking for.

According to the finalized descriptive figures and cleaned data sets, we choose the logistic regression model as our predictive approach to better predict our goal based on its high accuracy rate. It provides a measure of the size of the coefficient and a direction of the association. It also allows the measurement of several different types of variables by extension of the basic principles. We choose logistic regression because it can classify and evaluate the characteristics of each feature of a subject and calculate the practical level of each feature. The probability is the primary outcome that returns from the logistic regression model, and it could provide the comparison between these features and present the best coefficient. Lastly, the logistic regression model is the fastest and effective way to solve our business problem.

Inside the logistic regression model, the algorithm would return the calculated result through the sigmoid function back between 0 and 1. The coefficient number represents the credibility of classification results. In our data set, we choose 'Log GDP per capita', 'Social support', 'Healthy life expectancy at birth', 'Freedom to make life choices, 'Generosity', 'Perceptions of corruption', 'GNI', 'Total\_reserve' as our numeric variables and compare these variables to get the most effective one. Also, we choose the life ladder as our categorical variable and the dependent variable. After the data partition and the logistic regression with the penalty, we get our first feedback:

|  |  |
| --- | --- |
| Log GDP per capita | 0.948133 |
| Social support | 0.296414 |
| Healthy life expectancy at birth | 0.619485 |
| Freedom to make life choices | 0.208649 |
| Generosity | 0.140581 |
| Perceptions of corruption | - 0.005017 |
| GNI | - 0.242595 |
| Total\_reserve | 0.007173 |



**Table 2: Logistic Regression Model Output**

Then after we ran logistic regression with k-fold cross-validation with the value of 5, we specified within which range of the penalty levels we will search for the optimal penalty level and the level that leads to the best model candidate. We also further discretized the continuous alpha range [min\_alpha, max\_alpha] into n individual points of alpha and train n model candidates, each of which corresponds to one individual alpha point. Finally, we got a 0.9103 accuracy rate over the test partition.

**Results**

The analysis was carried out in two ways to get efficient and reliable output. We compared the results of each analysis for every variable. From the logistic regression results, Log GDP per capita had the highest coefficient. Log GDP per capita became the most vital factor that affected the happiness level of citizens in the country around the world. According to figure 2, a strong relationship was seen between the two variables giving a positive trendline analysis.

One of the interesting results that we came across was for the GNI index. Although both descriptive and predictive analysis indicates that there was a correlation between GNI index and Life Ladder, they gave exact opposite outputs. For the Logistic regression, GNI had a negative correlation which was contrary to the analysis discovered in appendix 2 of the descriptive analysis showing a positive correlation between the two variables. This may be indicative of Simpson’s Paradox, the phenomenon where an association between two variables emerges, disappears, or reverses when other variables are introduced into the model. When evaluating the simple correlation between GNI per Capita and Life Ladder, the correlation is positive in a subpopulation but reverses when examined again in the whole population within a Multiple Regression model. Because of this, we moved forward with the evaluation of the correlation shown in the Logistic Regression.

The health life expectancy variable had a high influence on the happiness level of the citizens in the country. For both Logistic regression and the descriptive chart, the same results were seen. There was a positive correlation between the two variables.

Another interesting finding was seen for the Total reserves variable. Contrary to our initial beliefs, it was seen that the total reserves of a country did not have much influence on the happiness level of a country. Both predictive and descriptive analysis suggested the same.

Based on logistic regression, a few other variables namely Freedom to make life choices, Social support, Generosity resulted in being high influencers to the happiness level. However, for the Perceptions of Corruptions variable, the absolute coefficient was the smallest which meant that it

had no significant correlation with the happiness level of citizens in their country.

**Conclusion**

The analysis reflects that financial indicators have an impact on the happiness level of a country. The correlation between the GDP per capita and the happiness ladder shows that a country’s happiness is influenced by its financial standing. While there are a couple of financial stability indicators like GNI and total reserves, it’s the GDP per capita that tends to influence the happiness ladder. Countries that are aiming to improve their happiness level should aim to improve their financial standing. These countries should start with introducing initiatives that work towards developing and growing the local economy, such as infrastructure and resources for industries and support to develop new industries like the digital economy. This would help improve the financial standing of a country by improving its GDP per capita which would lead to a happier country.

Similarly, if the country is aiming to improve its happiness level the impact of health indicators is also significant. Based on our findings an improvement in the healthy life expectancy at birth promotes the happiness level of a country. Policymakers in these countries should introduce programs that aim to improve the country's general health by working on the infrastructure and the health-related resources available to the people. This health improvement would improve life expectancy which would lead to a happier country.

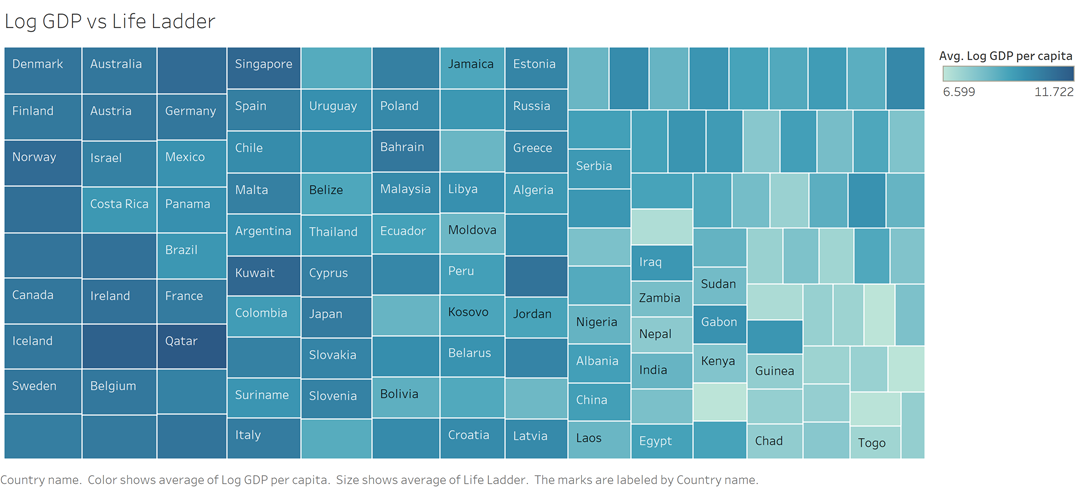
Our analysis was based on the available data, however, there are some limitations to the dataset and our analysis. The Happiness Ladder is based on a people survey that wasn’t collected by us. Due to this secondary collection of data, we can’t be sure as to how accurate these poll responses are. In addition to this, the data had many missing values that needed to be imputed and due to these missing values, a lot of data had to be taken out of our dataset.

**Future Study**

This report looks into the association of financial indicators and their correlation to the country's happiness level. While we did see a correlation with financial indicators, correlation with other indicators was also noticed. In the future, more other indicators such as health and social indicators should be analyzed, which would give a holistic overview of how a country's happiness level is impacted and what could be done to improve it. In addition to this, in the past year, Covid-19 has changed the world and with newer data, we could see how happiness patterns have changed over the past year and what specific indicators are impacting people’s happiness the most.

**Appendix**

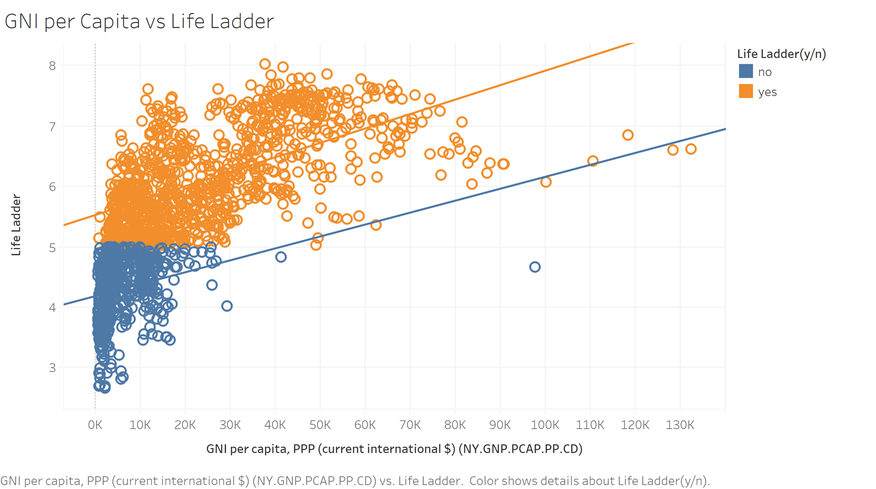
**Log GDP per Capita vs Life Ladder**



**Append 1: Tree map depicting Average Log GDP per Capita vs Average Life Ladder Score**

The treemap details the relationship between Average Log GDP per Capita versus Average Life Ladder score for each country over the course of the dataset timeline. Average Log GDP per Capita is represented by color, Box size represents Average Life Ladder score. The chart reinforces the positive relationship between the variables with size increasing and color intensifies from the bottom right to the top left. We concluded that increasing Log GDP per Capita will likely happen alongside increasing happiness reflected in the Life Ladder Score.

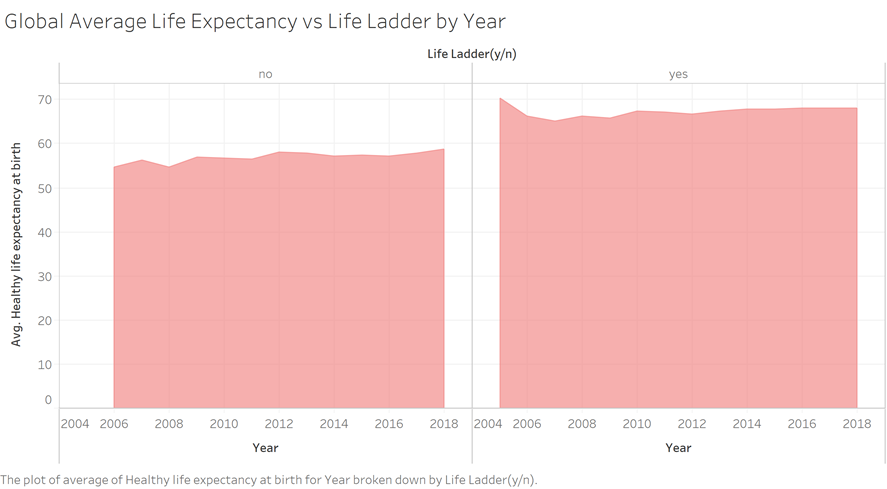
**GNI per Capita vs Life Ladder**



**Append 2: Scatterplot depicting GNI per Capita vs Life Ladder Score**

The scatter plot details GNI per Capita vs Life Ladder score with a differentiation filter between the “Yes” and “No” groups along with a trendline for both. The graph shows there is a positive correlation between GNI per capita and Life Ladder score in both groups with the relationship for the “Yes” group described as 2.38976e-05\*GNI per Capita + 5.52089 and the “No” group described as 1.97598e-05\*GNI per Capita + 4.17868. The “Yes” group has a stronger relationship that supports the association even more. For both groups, GNI per capita has good value as a predictor with a p-value of < 0.0001. The implication is that GNI per Capita has a strong correlation with the Life Ladder score and is a reliable predictor. Governments should focus greatly on improving the metric to promote happiness.

**Healthy Life Expectancy at Birth vs Life Ladder**



**Append 3: Area chart depicting Average Healthy Life Expectancy vs Life Ladder Grouping**

The Area chart details the Average Healthy Life Expectancy for the entire dataset overall years per the two Life Ladder groups. The “Yes” group which is Life Ladder Scores of > 0.5 has notably higher Average Healthy Life Expectancy. It appears to be a positive relationship. Average Healthy Life Expectancy has decent descriptive importance as per the graph. The graph details a steady average increase in year-to-year Average Healthy Life Expectancy. The insight gained further support higher Healthy Life Expectancy levels being associated with a happier score for a country.

**References**

[1] Herman, H. (2019). The pursuit of happiness. British Journal of Guidance & Counselling 2019, 47, 139-142

[2] Futrelle, D, (2020). Here’s how money can really buy you happiness. Times 2020

[3] Angrave, D., & Charlwood, A. (2015). What is the relationship between long working hours, over-employment, under-employment and the subjective well-being of workers? Longitudinal evidence from the UK. Human Relations, 68(9), 1491-1515.

[4] De Neve, J. E., & Oswald, A. J. (2012). Estimating the influence of life satisfaction and positive effect on later income using sibling fixed effects. Proceedings of the National Academy of Sciences, 109(49), 19953-19958.

[5] Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. Journal of economic psychology, 29(1), 94-122.

[6] De Neve, J. E., & Ward, G. (2017). Happiness at work. In the World Happiness Report 2017.

[7] Pearl, J, (2013). Understanding Simpson’s Paradox. UCLA 2013