**Final Report: World Happiness Report**

**Prajwal Narayanaswamy**

# Background

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

* 1. Questions
* What is the trend of Happiness over different locations?
* How did Happiness scores change between the 2015, 2016, 2017 and 2018?
* Is there a significant difference in Happiness Score over the years?
  1. Project Significance

The happiness scores and rankings use data from the Gallup World Poll. The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The scores are from nationally representative samples for the years 2015-2018 and use the Gallup weights to make the estimates representative. The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world’s lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

1. Data Analysis
   1. About the data

The raw data included 500 observations with 11 attributes. The data is complete without any missing values. Any missing values is replaced with a Zero to make the analysis feasible. The dataset contains entries of different variables which might contribute to the Happiness Score. COUNTRY is an attribute which contains the country names of all the recordings. GDP\_PER\_CAPITA is a decimal which contains GDP score across various observations. FAMILY refers to the score to which family plays a role in determining the Happiness Score. HEALTH refers to the extent to which Life expectancy contributed to the calculation of the Happiness Score. FREEDOM refers to the extent to which Freedom contributed to the calculation of the Happiness Score. GENEROSITY is a numerical value calculated based on poll participants' perceptions of generosity in their country. GOVERNMENT\_TRUST is the extent to which Perception of Corruption contributes to Happiness Score. DYSTOPIA\_RESIDUAL refers to a score based on a hypothetical comparison to the world's saddest country. YEAR refers to the time in which the observation was recorded. Finally, CONTINENT refers to the region of the country.

* 1. Data Processing

Any missing values in the dataset was replaced by Zero which could help better in the analysis of the data.

* 1. Understanding the Data

A general understanding of the data set can be obtained by looking at **descriptive statistics**. For example, year is more widely distributed with standard deviation of 1.08 and range of 3. The same applies to the happiness score with standard deviation of 1.09 and range of 4.93. The rest of the attributes have low Standard of Deviation and Range which indicates that they are less widely distributed.

Table1. Descriptive Summary of the Dataset.

Table, timeline

Description automatically generated with medium confidence

Table

Description automatically generated

It is essential to look that the Happiness Score Distribution occurring across the dataset regarding varying scores. It can be observed that the center of gravity for the scores lies in the range of 4.27~6.2

Figure1. Happiness Distribution

Chart, histogram

Description automatically generated

Further understanding of Happiness Score can be found plotting the score against the Year which gives an idea about the patter. With the help of **Tableau** software, a scatter plot was used to understand this question.

Figure3. Happiness vs Year

Chart

Description automatically generated

Now, we look at the variation of the Happiness Score with each variable which gives a clear understanding of the impact of the variables to the score.

Figure4. Happiness vs Different variablesA picture containing nature, image, day

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Based on the visualization. It is interesting to see that GDP per capita, family, freedom, dystopia residual has somewhat better linear relationship than generosity and government trust with the Happiness Score.

Furthermore, it is crucial to look closely at the geographic factor. The country and continents given in the dataset have been plotted on the map below with respect to the Happiness Score.

Figure5. Happiness Score vs Country

Map

Description automatically generated

Figure6. Happiness Score vs Continent

Chart

Description automatically generated

1. Methods
   1. Hypothesis

Foremost, we want to analyze the similarity of Happiness Score across different time. To determine this, we use the Hypothesis Testing Method.

**Null Hypothesis: There is similarity in happiness scores between 2015, 2016, 2017 and 2018  
Alternate Hypothesis: There is difference in happiness scores between 2015, 2016, 2017 and 2018**

Since there four different years, we conduct two separate two-tailed tests to observe the relationship. We must also ensure that the two-sample hypothesis testing satisfies all the requirements. The four samples are independent, with observations greater than 30. The level of significance is set to 0.05. The hypothesis testing is a two-tailed test

Text

Description automatically generated

* 1. Correlation and Regression
     1. Correlation

Before running a regression analysis, a basic correlation model was run. This is a preliminary step that we conduct in order to understand the data better and potentially spot the trends and relationships that may show up again in later models. We approach this by setting up a correlation matrix using the processed data and observing the values reported (-1<𝑥<1).**Jupyter Notebook was used to run the correlation**

* + 1. Regression

In order to further understand the data and the relationships between the given variables, we create a multiple linear regression model. We first construct a model including all variables and follow up by removing all variables with p-values above our level of significance. Finally, we run the regression model again with the remaining variables.

1. Results
   1. Hypothesis

T-Test method is used to determine the similarity of the Happiness Score. Optimizing the T-Test: Two-sample Assuming Unequal Variances, we obtained the following results:

Table2. Happiness Score for 2015 and 2016

Table

Description automatically generated

Since the P-Value between 2015 and 2016 is 0.98 which is greater than the assumed significance level (0.05), we can say that there is similarity between the Happiness Scores.

Table2. Happiness Score for 2016 and 2017 Table3. Happiness Score for 2017 and 2018

Table

Description automatically generated Table

Description automatically generated

The P-Value between 2016 and 2017 is 0.63 which is greater than the assumed significance level (0.05), we can say that there is similarity between the Happiness Scores.

The P-Value between 2017 and 2018 is 0.0046 which is less than the assumed significance level (0.05), we can say that there is a vast difference between the Happiness Scores.

The above results do not satisfy the Null Hypothesis. Hence based on Sample T-Testing with unequal variances, **we reject Null Hypothesis and conclude that there is difference in happiness score between 2015, 2016, 2017 and 2018.**

* 1. Correlation and Regression
     1. Correlation

To find the correlation matrix between the variables, Jupyter Notebook clubbed with Python Programming was used.

Figure5. Correlation Matrix

Graphical user interface, table

Description automatically generated

Figure7. Correlation Chart

Chart

Description automatically generated with medium confidence

* + 1. Regression

The data contains a combination of Categorical and Quantitative variables. To obtain good regression results, it is important to consider both quantitative and categorical variables in the model.

* + - 1. Quantitative variables

Quantitative variables are attributes which contain only numbers as Data. Except Country and Continent, our data is entirely comprised of Quantitative Variables.

Table4. Example snippet of Quantitative Data

Table

Description automatically generated

After running a regression analysis, we get an Adjusted R square value of 0.78 which indicates that these variables account to 78% of Happiness Score. The significance F value is 5.03E-162 which is far lower than 0.05. Hence this model can be considered. Further, it can be observed that apart from the variable Government Trust, the P-Value for all other variables is very much lesser than 0.05 which indicates that these variables play a role in calculating Happiness Score.

Table 5. Regression\_Quant

Table

Description automatically generated

With this analysis, the regression equation is

***Happiness Score = 2.072008 + (1.6173\* gdp\_per\_capita) -(0.7147\*family) + (1.1806\*Health) + (2.0052\*Freedom) + (0.7642\*Generosity) + (0.4520\*Dystopia\_Residual)***

* + - 1. Categorical Variables

The categorical variables in our data are Country and Continent. Since we are analyzing data for only 2015-2018, I have considered Year also as a category. To ease the process, the Continent variable was converted to numeric sequence from 1 to 6.

Table6. Example snippet of Categorical Data converted to numbers

***Table

Description automatically generated***

After running a regression analysis, we get an Adjusted R square value of 0.28 which indicates that these variables account to 28% of Happiness Score. The significance F value is 5.619E-38 which is far lower than 0.05. Hence this model can be considered. Further, it can be observed that the Year variable has a P-Value greater than 0.05, which implies that it does not play any major role in decision of Happiness Score.

Table7. Regression\_Categorical

Table

Description automatically generated

With this analysis, the regression equation is

**Happiness Score = -182.335 +(0.0937\*Year) -(0.2964 \* Continent\_Categ)**

* + - 1. Quantitative + Categorical Variables

To obtain the best regression mode, it is always a good practice to combine both quantitative and categorical data and observe the results.

Table8. Example Snippet of Quant + Categorical Data

Table

Description automatically generated

After running a regression analysis, we get an Adjusted R square value of 0.8006 which indicates that these variables account to 80.06% of Happiness Score. The significance F value is 2.92E-167 which is far lower than 0.05. Hence this model can be considered.

Apart from this many iterations of analysis were done to determine the best R square value. But it turned out that including Continent and Year yielded the best results.

Table9. Regression Quant + Categorical

Table

Description automatically generated

With this analysis, the regression equation is

**Happiness\_Score = -536.5274 + (1.5324 \* gdp\_per\_capita) - (0.3102 \* Family) + (1.0766 \* Health) + (1.6950\*Freedom) + (0.8595 \* Generosity) + (0.5743 \* Government\_Trust) +(0.4651 \* dystopia\_residual) + (0.2671\*Year)**

**Validation**

From all the above Regression Analysis, it can be seen that the Quantitative + Categorical Analysis yielded the best result with an Adjusted R square value of 0.8006. To verify this, a validation was done by dividing the data into training and testing. The training dataset included around 80% (400 observations) of the dataset and testing data included 20% (100 observations) of the dataset.

The regression analysis was once again done on the training dataset, but now with both Categorical and Quantitative values. The HS Cap is the calculated Happiness Score after substituting all the values into the regression model.

Table10. Training Set Regression

Table

Description automatically generated

The HS Cap is the calculated Happiness Score after substituting all the values into the regression model equation. And the Error is the absolute value of the difference of the Actual Score to Calculated Score.

Table 11

Table

Description automatically generated

**We get an Error Difference of 4.314% between the training dataset and validation dataset which implies our model is efficient.**

Plotting Happiness Score vs HS Cap to visualize the model, we get

Figure 8

1. Conclusions

This project investigates multiple aspects of the parameters contributing to the calculation of Happiness Score. Through the data processing and the analysis performed, the data set entries were sorted out, and a visualization of the trend line was performed.

Through hypothesis testing, we were able to verify that there is significant difference in the Happiness Score over the time. To perform forecasting analysis, the dataset lacked proper time information to perform the time-series analysis which resulted in not choosing the forecasting method to analyze the data.

Construction of multiple regression models lead to understand the relationship between the various parameters contributing to the calculation of Happiness Score. As previously mentioned, we were able to construct a model of an adjusted R Square of 0.8006 which indicated that we were able to account for 80.06% of the influence of parameters on the Happiness Score. This is well above the threshold of 0.75 which indicates that our model works well with the chosen dataset.

1. Future Work

Another approach for the future work is deep diving into data with usage of machine learning models to get more accurate regression models. There are many data which are not recorded and is filled with 0. Filling these values might affect the overall accuracy of the model and this can be explored with full range dataset.