Selma Sentissi El Idrissi

selmasentissi30@gmail.com

Happiness

Predicting and Understanding Happiness Score in Various Countries

Table of Contents

[Introduction](#_Introduction)

[Methodology](#_Methodology)

[Data Cleaning](#_Data_Presentation_and)

[Exploratory Data Analysis](#_Exploratory_Data_Analysis)

* [*Individual Years*](#_Individual_Years)
* [*Final Dataset*](#_Final_Dataset)

[Modeling](#_Modeling)

* [*Models Used*](#_Models_Used)
* [*Results*](#_Results)
* [*Feature Importance Methods*](#_Feature_Importance_Methods)

[Dashboard](#_Dashboard)

[Conclusion](#_Conclusion)

# Introduction

In this project, the main goal is to understand which variables impact people’s happiness. The Happiness Score can be defined as the national average of the responses to the main life evaluation question asked to citizens. Bhutan is a small country in Asia and it introduced the Gross National Happiness (GNP) to the world [[1]](#footnote-1). Bhutan is also a developing country with a low income economy which makes you think that a country’s Gross Domestic Product (GDP) is not the only factor that creates happiness within a civilization. This research will determine what makes the citizens of a country happy, and on the opposite side, what contributes to their unsatisfaction. If money does not buy happiness, what does?

I will use a dataset that can be found [here](https://www.kaggle.com/mathurinache/world-happiness-report?select=2016.csv) which includes 6 years of data from around 150 countries per year. Each dataset has variables that affect the GHP such as social support, corruption perception… I will first find correlations between these variables and the countries’ geographic locations, then I will also find correlations at various scales (continental, regional…), and finally, I will apply various models to try and predict future year’s happiness scores. I will be producing a summary report, as well as a notebook, and a [dashboard](https://public.tableau.com/app/profile/selma.sentissi/viz/Happiness_Story_16291295512870/Story1?publish=yes) to summarize my findings. All of the above can be found [here](https://github.com/selmasentissi-design/Happiness).

At the conclusion of this project, I expect to find that GDP does have an important impact on happiness, however, that other important variables can be as impacting, in the positive or negative direction.

# Methodology



# Data Presentation and Cleaning

*2015:*

The data for the year 2015 has no null values so no cleaning was needed for individual Exploratory Data Analysis (EDA) and modeling. The columns used were Country, Region, Happiness Rank, Happiness, Standard Error, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, and Dystopia Residual.

*2016:*

The data for the year 2016 also has no null values so no cleaning was needed for individual Exploratory Data Analysis (EDA). The columns used were Country, Region, Happiness Rank, Happiness, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, and Dystopia Residual. It has the same variables than the 2015 dataset without the ‘Standard Error’ column.

*2017:*

The data for the year 2017 also has no null values so no cleaning was needed for individual Exploratory Data Analysis (EDA). The columns used were Country, Happiness Rank, Happiness, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, Upper Confidence Interval, Lower Confidence Interval, and Dystopia Residual. The dataset does not have the Standard Error and the Region column that the 2015 dataset has but has an additional two columns: Upper Confidence Interval and Lower Confidence Interval.

*2018:*

The data for the year 2018 also has no null values so no cleaning was needed for individual Exploratory Data Analysis (EDA). The columns used were Country, Happiness Rank, Happiness, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, and Dystopia Residual. This dataset does not have the rows: Region, Upper Confidence Interval, Lower Confidence Interval, and Dystopia Residual.

*2019:*

The data for the year 2019 also has no null values so no cleaning was needed for individual Exploratory Data Analysis (EDA). The columns used were Country, Happiness Rank, Happiness, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, and Dystopia Residual. This dataset does not have the rows: Region, Upper Confidence Interval, Lower Confidence Interval, and Dystopia Residual.

2020:

The data for the year 2020 initially had the most variables and had no null values. On top of the initial variables that the data for the year of 2015 has, it also has ‘Explained by: Perceptions of corruption', 'Explained by: Log GDP per capita', 'Explained by: Generosity', 'Explained by: Freedom to make life choices', 'Explained by: Healthy life expectancy', 'Explained by: Social support', 'Region', 'Standard Error', 'Upper Confidence Interval', 'Lower Confidence Interval', and 'Ladder score in Dystopia'.

# Exploratory Data Analysis

The Exploratory Data Analysis is split in two parts: first, I focused on each year on their own to draw general analysis and eventually find a pattern along the years; second, the analysis is focused on the final dataset.

## Individual Years

For each year, I computed the top 10 happiest countries in descending order (from the most happy to the least happy), and when applicable (when the data had the variable ‘Region’), the regions from which the 50 happiest countries are from.

*Top 10 Happiest Countries by Year:*

* In **2015**, Switzerland, Iceland, and Denmark are the three countries with the highest Happiness Score with scores of 7.587, 7.561, and 7.527. These three countries are part of Western Europe. Also, within the 50 countries that ranked the highest in the Happiness Score, most of them are located in North America, Australia and New Zealand, followed by Western Europe.
* In **2016**, Switzerland, Iceland, and Denmark are still the three countries with the highest Happiness Score but in a different order. This time, Denmark has a higher happiness score than Switzerland with a with scores of 7.526 for Denmark, 7.509 for Switzerland, and 7.501 for Iceland. These three countries are part of Western Europe. Furthermore, within the 50 countries that ranked the highest in the Happiness Score, most of them are located in Australia and new Zealand, North America, followed by Western Europe.
* In **2017,** Switzerland lost its place in the top 3 happiest countries and has been replaced by Norway which went straight to the first position with a score of 7.537, followed by Denmark and Iceland with respective scores of 7.522 and 7.504.
* Once again, a country that was not in the top 3 went straight to first place. Finland is the happiest country in **2018** with a happiness score if 7.632 followed by Norway and Denmark with respective scores of 7.594 and 7.555.
* Norway lost another place in **2019** to Denmark and Finland kept its first place with a score of 7.769 and Denmark and Norway with scores of 7.600 and 7.554.
* In **2020,** Norway lost 3 places and is now the 5th happiest country and Finland kept its first place with a score of 7.808 and is still followed by Denmark but now also followed by Switzerland with respective scores of 7.646 and 7.560.

Overall, the 5 countries with the highest happiness score on average over the years are Finland, Denmark, Norway, Iceland, and Switzerland.

## Final Dataset

In the final dataset, we used the columns that were in all years’ datasets. So the final dataset has the columns: Country, happiness Rank, Happiness Score, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, and Year.

Following an initial correlation feature heatmap, we can see that nearly all variables are strongly related with the Happiness Score. Then, I grouped the countries by their happiness rank to see which ones were the happiest on average. Denmark is the country with the happiest people on average with an average happiness rank of 2.17, followed equally by Finland and Norway which both have a score of 3.17. On the other hand, the country with the least happy people on average seems to be Syria with an average rank of 152.6, followed by Central African Republic and Rwanda with respective scores of 152.40 and 151.67.

The data suggests also that Angola, Sudan, and Syria are the three countries with the lowest Freedom score over the years on average with respective scores of .027, .033, and .082 whereas the Maldives, Hong Kong, and Uzbekistan are the three countries with the highest Freedom score over the years on average with respective scores of .854, .780, and .710. Furthermore, Somalia, Djibouti, and Sudan are the three countries with the lowest GDP per Capita on average over the years with respective scores of .005, .440, and , 604. Hong Kong, Maldives, and Taiwan are the three countries with the highest GDP per Capita on average over the years with respective scores of 10.935, 9.519, and 6.105. All the countries with the lowest Freedom and GDP score belong to the continent of Africa whereas all the countries with the highest scores in these variables are from Asia.

# Modeling

I computed Machine Learning Models on the final dataset to identify patterns and eventually draw conclusions on which variables impact the most (or the least) peoples’ happiness within a country. I also used feature importance methods to determine which variables are the most important when predicting happiness.

## Models Used

Initially, I used a linear regression model as a baseline. Then, I used Random Forest Regressor which is a more flexible model. I used feature importance methods with both Machine Learnings models. In addition, I also tried the Kernel Ridge Model which offers a slightly more flexible approach than the Linear Regression. The last model I computed is the Boosting model which is an ensemble method made of multiple algorithms. All the models have a train ratio that is 75% of the final dataset, a 15% for the validation ratio, and a test ratio of 10%.

## 

## Results

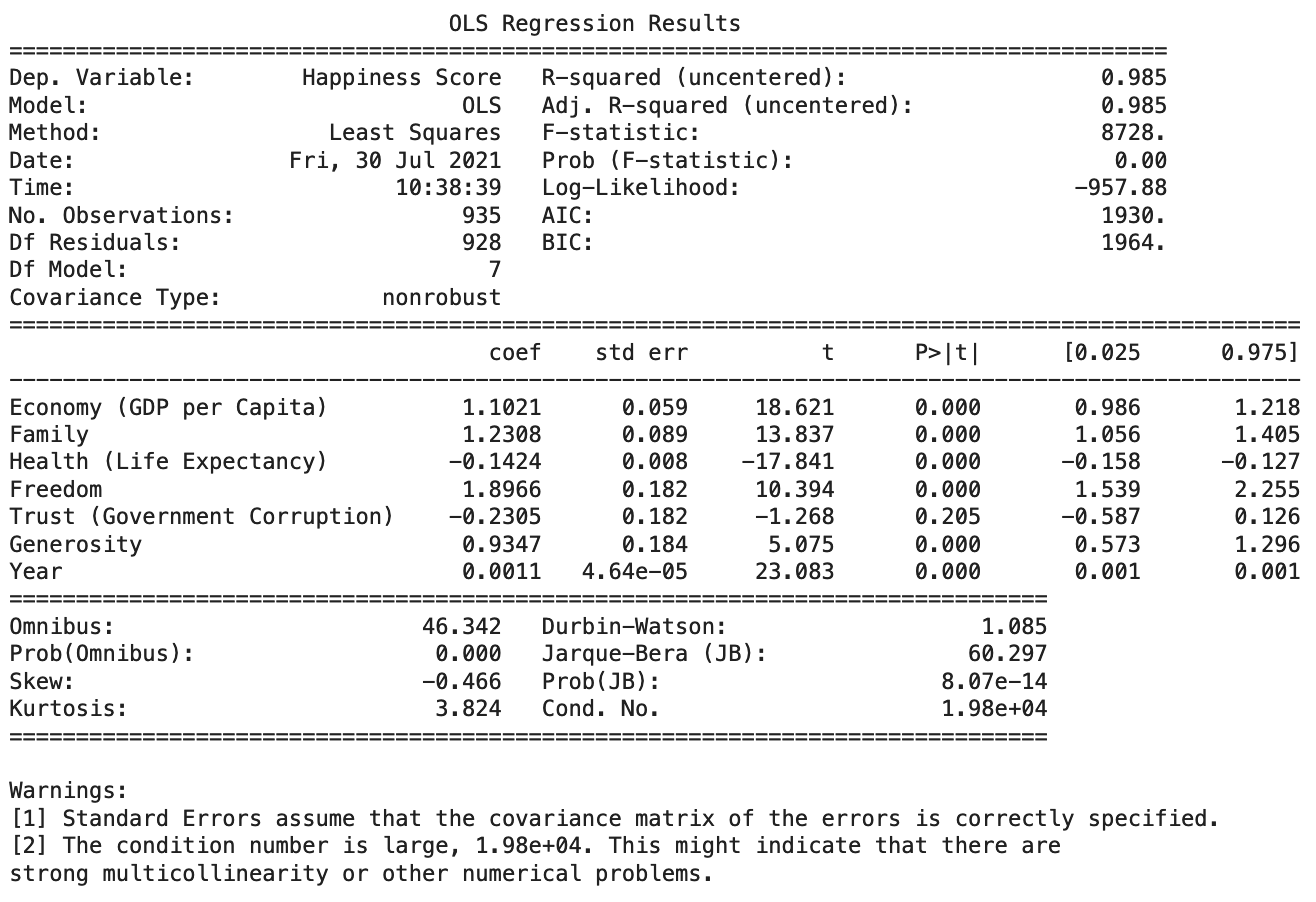
Here is a summary of the R squared values for all the models used for this analysis:

## Table Description automatically generated

The Random Forest model is the most accurate with a R squared value of .76, followed by the Kernel Ridge model with a R squared value of .62. The Boosting model is not accurate at all with a R square of .0001 and a negative R squared value for both the R squared value for the training dataset and the validation dataset.

## Feature Importance Methods

*Ordinary Least Squares:*

**

The first feature importance used was the Ordinary Least Squares method (OLS). This method suggested that all variables were significant with a P-value of .000 except for ‘Trust (Government Corruption)’ which had a P-value of .205. The ones with the highest coefficient amongst these variables are: Freedom with a coefficient of 1.897, Family with a coefficient of 1.231, and the Economy (GDP per Capita) with a coefficient of 1.102.

*Random Forest Feature Importance:*

*Chart, bar chart

Description automatically generated*

The second method used was the Random Forest Feature Importance. In this method, the ‘Health (Life Expectancy)’ variable is the most significant with a magnitude of .35 followed by the ‘Economy (GDP per Capita)’ with a magnitude of .34 which is then followed by the ‘Family’ variable with a .30 magnitude. The variable that has the lowest magnitude- except for the ‘Year’ variable- (<.08) is the ‘Trust (Government Corruption)’ which confirms the OLS method above.

*Shapley Additive exPlanations:*

*A picture containing chart

Description automatically generated*

The last method used was the Shapley Additive exPlanations (SHAP). In this case, the ‘Economy (GDP per Capita)’ is the feature with the highest importance amongst the one in the dataset. In this case again, the least important feature is the ‘Trust (Government Corruption)’.

# Dashboard

*Link to the Dashboard Story:* <https://public.tableau.com/app/profile/selma.sentissi/viz/Happiness_Story_16291295512870/Story1?publish=yes>

In the first Dashboard Sheet of the story ‘World Happiness’, we can see the average of countries’ happiness scores over the years. We can clearly see darker tones in Western Europe as well as Northern America which means that people have highest happiness scores in these places.

In the second Dashboard Sheet ‘Top Highest GDP’, we can see the top 10 countries with the highest GDP per capita. I used this variable since we could see in the feature importance methods that this variable was always in the top 3.

In the third Dashboard Sheet ‘Top Highest Life Expectancy’, there are the top 15 countries with the highest life expectancy, which is another one of the top 3 most important features for the random forest feature importance methods.

In the fourth dashboard ‘Top Highest Family’, there are the 10 countries with the highest score of Family over the years. Family is the second most important feature in the OLS method and the third most important feature in the other methods.

Finally, in the fifth dashboard ‘Top Highest Freedom’, we can see the countries with the highest score of Freedom over the years. Freedom was the most important feature in the OLS regression as well as the fourth most important in the SHAP and Random Forest methods.

# Conclusion

This research project about happiness includes Exploratory Data Analysis on each year of data from 2015 to 2020 separately, as well as on a final dataset that combines all years. It also includes models to determine feature importance within the variables studied to figure out which variables are significant and which are not significant. The four most important features amongst the different feature importance models are: ‘Economy (GDP per Capita)’, ‘Health (Life Expectancy)’, ‘Family’, and ‘Freedom’. This project also includes a story dashboard on the four main variables to visualize them and illustrate the findings stated above. Thus, the hypothesis stated in the introduction is valid: GDP does play an important part in a country’s citizens’ happiness but it is not the only impactful variable.

1. https://sustainabledevelopment.un.org/index.php?page=view&type=99&nr=266&menu=1449 [↑](#footnote-ref-1)