Bahareh Zohouri

Analyzing World happiness scores and countries' future rank prediction

1. Introduction

1.1. Background

The world happiness report (WHR), positions worldwide nations by their bliss levels. The first copy of this report was published in 2012 and until 2020 eight reports have been published yearly. The main information for measuring joy and happiness comes from the Gallup World Survey. In this survey, the respondents are asked to rate their current lives from 0 to 10, in which 0 corresponds to the most exceedingly bad conceivable life and 10 identifies with the most excellent conceivable life for them.

The final Happiness Score is a national average of the responses to the main life evaluation question asked in the Gallup World Poll (GWP), which uses the Cantril Ladder.

This report contains the Happiness Score for 153 countries. The Happiness Score is explained by the following factors:

- GDP per capita
- Healthy Life Expectancy
- Social support
- Freedom to make life choices
- Generosity
- Corruption Perception
- Residual error

The WHR explores the perception of these nations' residents and will help governments, organizations, and respectful society to educate their policy-making choices. Moreover, it will reveal the success model of the countries such as Denmark so that other countries may be able to learn and follow to improve their nations' wellbeing.

1.2. Problem statement

Almost all policymakers search for solutions to improve the quality of life of the residents. To reach this goal, it is essential to explore the underlying factors, their interdependencies, their mutual effects, previous trends and progress and more importantly recognizing the areas which the country can invest in to improve this happiness score. Without a careful study of multiple determinant variables, it would be difficult for countries to initiate a development plan. Imagine that the Canadian government has planned to acquire the first rank in this report for 2021. How they can accomplish this objective and what are the main determinants?

Therefore, this study aims to study different independent variables such as geographical region, income etc. by categorizing them and how they may affect the final happiness score. Moreover, previous trends would be explored both for variables and countries and finally building a model to predict the score and rank of Canada in 2021.

1.3. Stakeholders

This analysis will interest various stakeholders. Readers may be drawn in by wanting to know how their nation is performing. Moreover, they will be able to understand how different variables will correlate to a better life in the happiest countries.

This analytical report will be also of interest to governmental parties, research groups and policymakers since it will reveal the past trends and the effectiveness of their previous development strategies. Besides, they would be able to project their country's rank in the coming years and which factors they should work on more seriously to reach that goal.

2. Data

2.1. Data source

The information related to the World Happiness Report rank has been acquired in the excel format from Kaggle website. However, the main source of data is Gallup and different reports have been generated by different research teams who were partnered with this institution. There are six separated excel files which contain information regarding countries' score and rank as well as their score in each different variable from 2015 to 2020.

2.2. Data pre-processing

At first sight, it could be observed that the number of countries under study vary over different years. Therefore, I conducted a VLOOKUP function in excel to find what countries are missing in each file. The total number of countries and missing countries are shown in the table below:

Table 1.Data pre-processing overview

year	Total number of countries	Compared with	Number of missing countries	Name of missing countries
2015	158	2016	5	Namibia, South Sudan, Belize, Somalia, Purto Rico
		2017	6	Belize, Taiwan, Hong Kong, Somalia, Namibia, South Sudan
		2018	6	Trinidad and Tobago, Belize, North Cyprus, Somalia, Namibia, South Sudan
		2019	7	Trinidad and Tobago, North Cyprus, North Macedonia, Somalia, Namibia, Gambia, South Sudan
		2020	6	Taiwan, Hong Kong, Maldives, Gambia, Namibia, South Sudan
2016	157	2015	3	Oman, Djibouti, Central Africa
		2017	5	Taiwan, Hong Kong, Mozambique, Lesotho, Central African Republic

		2018	5	Trinidad and Tobago, North Cyprus, Mozambique, Lesotho, Central African Republic
		2019	8	Trinidad and Tobago, North Cyprus, North Macedonia, Gambia, Mozambique, Swaziland, Lesotho, Central African Republic
		2020	8	Taiwan, Congo, Hong Kong, Gambia, Mozambique, Swaziland, Lesotho, Central African Republic
2017	155	2015	9	Oman, Djibouti, Swaziland, Taiwan, Suriname, Hong kong, Somaliland region, Laos, Comoros
		2016	7	Purto Rico, Taiwan, Suriname, Hong kong, Somaliland region, Laos, Comoros
		2018	5	Taiwan, Trinidad and Tobago, North Cyprus, Hong Kong, Laos
		2019	9	Taiwan, Trinidad and Tobago, North Cyprus, Hong Kong, Laos, North Macedonia, Gambia, Swaziland, Comoros
		2020	6	Hong Kong, Maldives, Laos, Gambia, Swaziland, Comoros
2018	156	2015	8	Oman, Suriname, Trinidad and Tobago, North Cyprus, Swaziland, Somaliland region, Comoros, Djibouti
		2016	6	Purto Rico, Suriname, Trinidad and Tobago, North Cyprus, Somaliland region, Comoros
		2017	4	Taiwan, Trinidad and Tobago, North Cyprus, Hong Kong
		2019	4	North Macedonia, Gambia, Swaziland, Comoros
		2020	8	Taiwan, Trinidad and Tobago, North Cyprus, Hong Kong, Maldives, Gambia, Swaziland, Comoros
2019	156	2015	9	Oman, Suriname, Trinidad and Tobago, North Cyprus, Somaliland region, Macedonia, Sudan, Djibouti, Angola
		2016	9	Purto Rico, Suriname, Trinidad and Tobago, Belize, North Cyprus, Macedonia, Somaliland region, Sudan, Angola

		2017	8	Taiwan, Trinidad and Tobago, Belize, North Cyprus, Macedonia, Hong Kong, Sudan, Angola
		2018		Belize, Macedonia, Sudan, Angola
		2020	10	Taiwan, Trinidad and Tobago, North Cyprus, Hong Kong, Maldives, Macedonia
2020	153	2015	11	Oman, Qatar, Taiwan, Suriname, Hong Kong, Bhutan, Somaliland region, Sudan, Djibouti, Angola, Syria
		2016	11	Purto Rico, Suriname, Qatar, Taiwan, Belize, Hong Kong, Somalia, Somaliland region, Sudan, Angola, Syria
		2017	8	Qatar, Belize, Hong Kong, Somalia, Bhutan, Sudan, Angola, Syria
		2018	11	Taiwan, Qatar, Trinidad and Tobago, Belize, Hong Kong, North Cyprus, Bhutan, Somalia, Sudan, Angola, Syria
		2019	9	Taiwan, Qatar, Trinidad and Tobago, North Cyprus, Hong Kong, Bhutan, Somalia, North Macedonia, Syria

Since I do not have the information for the missing countries to complete the report of 2020, I dropped the rows related to these countries while data cleaning.

Moreover, some features include redundant information and should have been dropped. In the next sections, an Exploratory data analysis including correlations, Descriptive analysis, etc. will be done to prepare data for further analysis. And finally, the prediction model will be suggested.

2.3. Data wrangling

After importing the datasets, I checked the datasets for any missing values. Other than the file pertained to the year 2018, all the cells were complete. Therefore, I replaced the missing data in 2018 with NaN which is defined by Python. Then we will replace this non-available number with the average of corresponding cells in other years (0.368386).

In order to compare and merge the data from different years, it is necessary to have similar columns in each years' data frame. Therefore, after discovering the name of the columns, we will drop unnecessary columns and rename the rest so that they have a similar column name. This is called "Data formatting".

Features of the data frame in all years after dropping unnecessary columns and renaming them are as follows:

Table 2.Feature selection

Country	Region	Happiness Rank	Happiness Score
Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom
Trust (Government Corruption)	Generosity		

Also, for the year 2020, there is no Happiness rank which is added to the dataset by inserting a column with row numbers.

3. Methodology

This section will be divided in two parts. In the first part, I employed an Exploratory Data Analysis to reveal the structure of the data. I used correlation, grouping and various data visualization tools such as maps and scatter plots. In the second part, I tried to fit a model to predict Happiness Score of the countries.

3.1. Exploratory Data Analysis

3.1.1. Correlation between Independent features and Happiness score

In this section, I want to know how each of independent variable such as GDP per Capita, Family, Health(life expectancy), freedom, Generosity and Trust (Government Corruption) are correlated in a country's happiness rank and which of them has greater importance in defining this score.

As can be seen in the following figures, all variables including GDP per Capita, Family, Health (life expectancy), freedom, Generosity except than Trust in Government (in terms of Corruption) have positive correlation with Happiness Score. Not surprisingly, we cannot conclude that countries with higher GDP per Capita or more freedom are happier than other countries. In terms of Trust in government, as can be observed in figure.6, in the countries where people believed in higher governmental corruption, they reported to be less happy. However, low correlation (0.3943) between this variable and Happiness score reveals that this feature can not be a good predictor for happiness. Moreover, regarding the Generosity variable, it does not seem like a good predictor of the happiness at all since the regression line is close to horizontal.

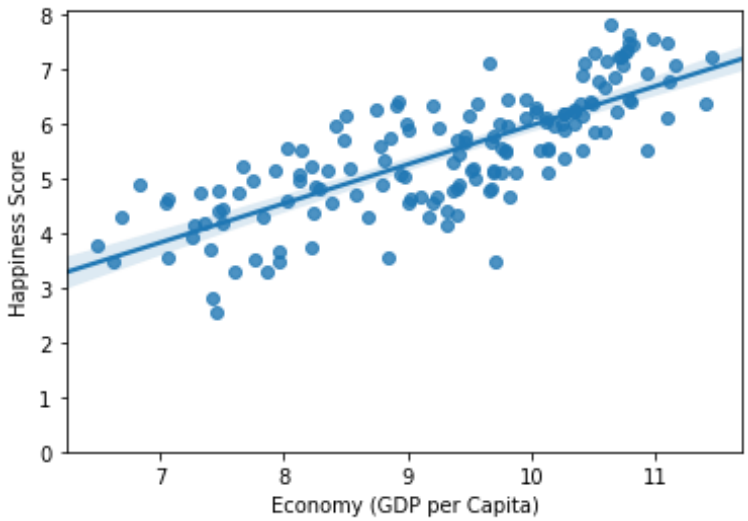

Figure 1.Relationship between Economy and Happiness Score

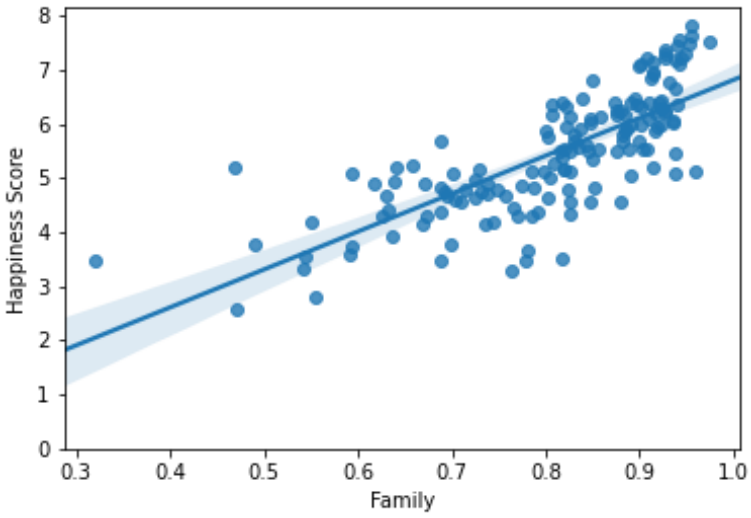

Figure 2.Relationship between Family and Happiness Score

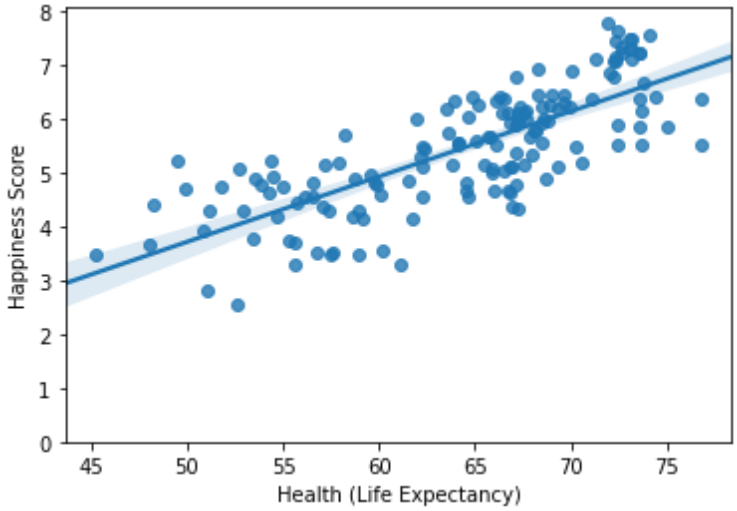

Figure 3.Relationship between Health and Happiness Score

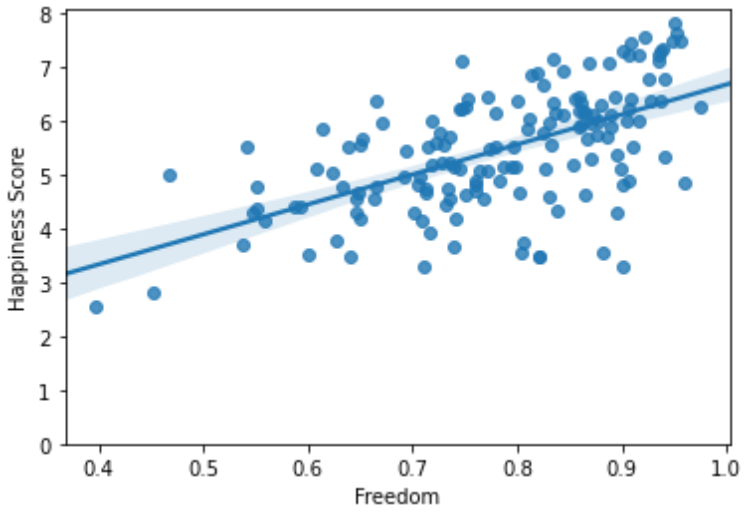

Figure 4.Relationship between Freedom and Happiness Score

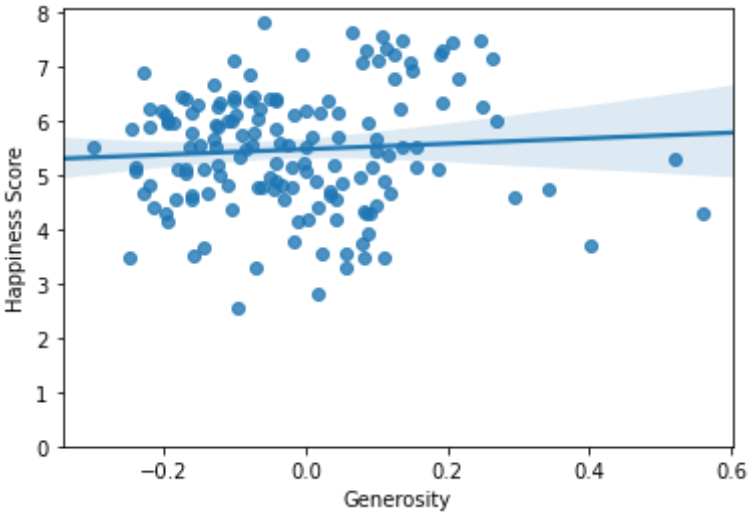

Figure 5.Relationship between Generosity and Happiness Score

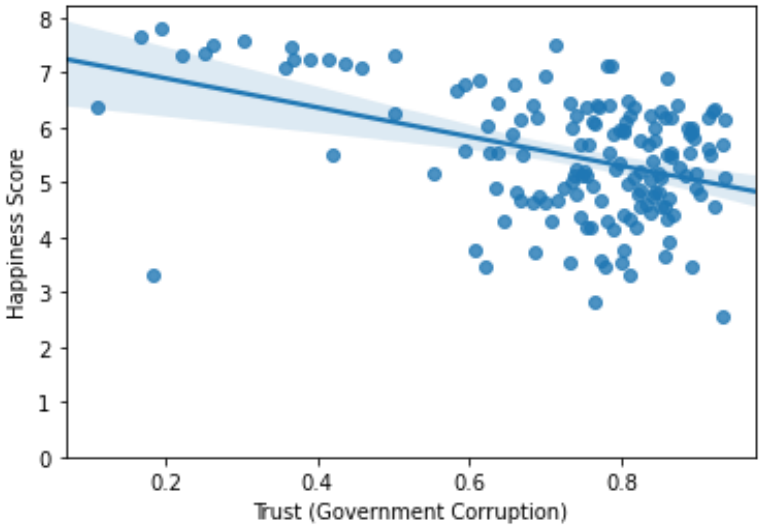

Figure 6.Relationship between Government corruption and Happiness Score

3.1.2. Happiness in different regions

In this section, we want to know what is the average of happiness in each region of the world, in 2015 and 2016 for which the information about the region is available. To reach this objective, I used the Gropby() function and the countries were grouped by their region in ten regions as follows.

Table 3. World happiness dispersion in 2015

	Region	Happiness Score
0	Australia and New Zealand	7.285000
1	Central and Eastern Europe	5.332931
2	Eastern Asia	5.626167
3	Latin America and Caribbean	6.144682
4	Middle East and Northern Africa	5.406900
5	North America	7.273000
6	Southeastern Asia	5.317444
7	Southern Asia	4.580857
8	Sub-Saharan Africa	4.202800
9	Western Europe	6.689619

Table 4.World happiness dispersion in 2016

	Region	Happiness Score
0	Australia and New Zealand	7.323500
1	Central and Eastern Europe	5.370690
2	Eastern Asia	5.624167
3	Latin America and Caribbean	6.101750
4	Middle East and Northern Africa	5.386053
5	North America	7.254000
6	Southeastern Asia	5.338889
7	Southern Asia	4.563286
8	Sub-Saharan Africa	4.136421
9	Western Europe	6.685667

When Plotting the bar chart for 2015, it is concluded that the average Happiness Score of Australia and New Zealand the highest in the world followed by North America and Western Europe. This is shown in the following figure:

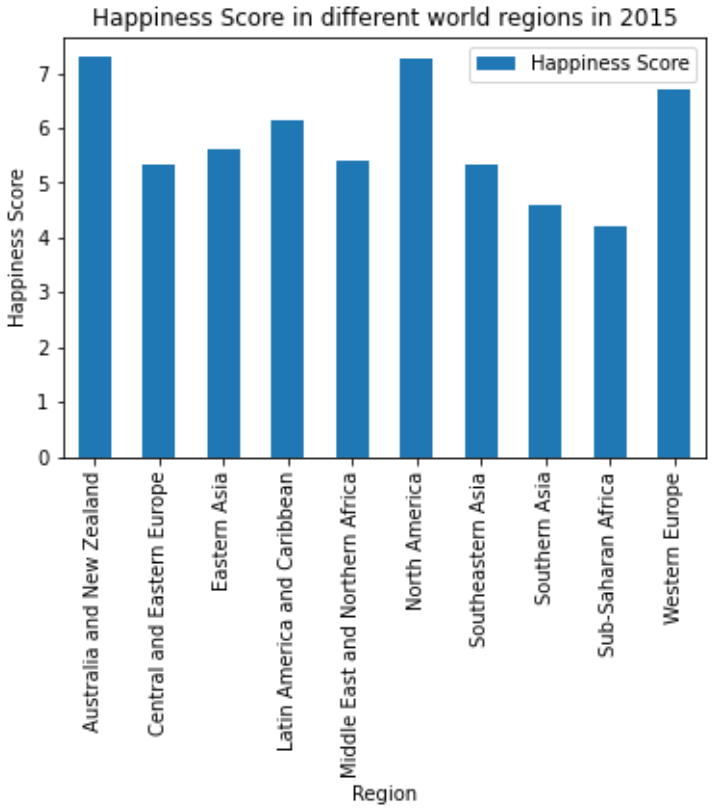

Figure 7.Happiness Score in different world regions in 2015

3.1.3. Regions with most Unhappy countries

Based on the following table, among the 20 unhappiest countries, 15 countries belong to Sub-Saharan Africa, 2 countries from the Middle East and North Africa region, 2 South Asian countries and 1 country from Latin America and the Caribbean region.

Region	Country
Sub-Saharan Africa	15
Middle East and North Africa	2
South Asia	2
Latin America and Caribbean	1

Figure 8.Most unhappy regions

3.1.4. Choropleth Maps

Here, I want to show the happiest countries in 2020 on the world map. The darker is the colour of a country on the map, the happier it is. As can be seen, North America, Europe and Australia have the most happiest countries.

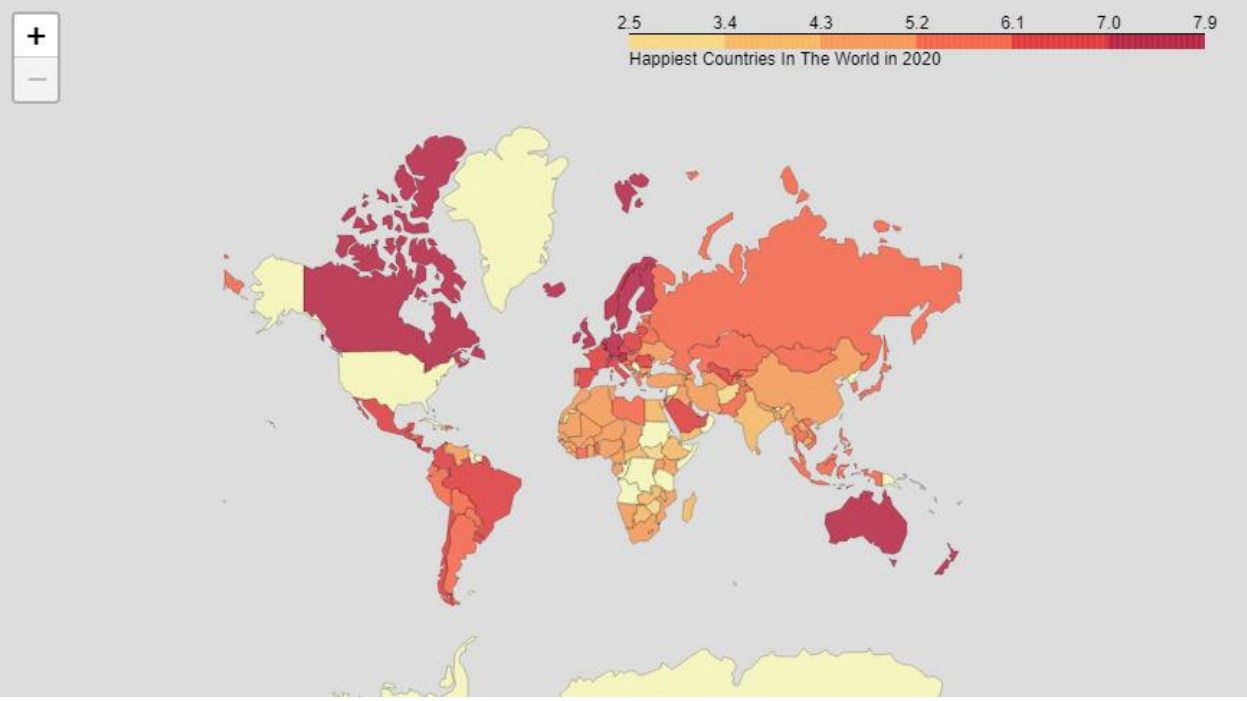

Figure 9. Map of world happiness score

3.2. Model development

3.2.1. Multiple Linear regression

In this section, I should discover how different variables will help to predict the Happiness score and a countries rank.

Since multiple variables predict the Happiness Score of a country, I used multiple linear regression. After training the model and calculating the coefficient of different variables including Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Generosity and Trust (Government Corruption), the coefficient would be as follows:

Coefficients: [[0.20699327 2.84803982 0.03594441 1.79365182 0.53553804 -0.37503036]]

Later, I used the Ordinary Least Squares method to solve the problem. When calculating the residual sum of square and Variance score, they respectively equalled to 0.22 and 0.8 which is not pleasant. I repeated this procedure by dropping the "Generosity" variable which had a weak correlation with happiness score and these numbers changed to 0.21 and 0.81. This change shows that ignoring this variable in our prediction model could improve our model quality.

3.2.2. Model building using Sklearn

For building a model based on our observations, I used Sklearn. For this purpose, first I will fit the model using all independent variables including Generosity. The result is as follows:

Table 5.Model outputs with Generosity

R^2 score for train	0.7726096772336093
R^2 score for test	0.5782217120641335
mean squared error train	0.30241766040521667
mean squared error test	0.3496565987103075

Then I will fit the model with all the variables except than "Generosity" and the result would be as follows:

Table 6.Model outputs excluding Generosity

R^2 score for train	0.7691473959749999
R^2 score for test	0.5821926013248095
mean squared error train	0.30702231985226436
mean squared error test	0.34636470893680077

It can be observed that this change did not help to decrease the mean squared error to a great extent, however, the numbers are low for both practices and still, other algorithms may improve the performance.

4. Result and discussion

Studying different variables revealed that economic situation, family support and life expectancy have the highest impact that a country resident feels happy. Freedom to make life choices has less impact and there is no meaningful correlation between generosity and happiness score in countries. In other words, whether a person has donated to a charity in past months is not a determinant of his happiness.

Reviewing the determinant variable also shows that in countries with higher corruption perception, the happiness score is lower. In other words, these two variables have a negative correlation with each other.

The happiest countries in the world are mainly situated in Australia and New Zealand, North America and western Europe.

In terms of less happy countries, most of them are in Sub-Saharan Africa.

The happiness score of Finland has been increasing during the past years and has the highest score since 2018.

Among the countries under study, the happiness score of India is decreasing over time compared to other countries where the trend is constant or increasing.

The ten healthiest countries in the world in 2020 are all western European countries except New Zealand which places at 6th rank.

When it comes to prediction and modeling, it can be inferred that a model which is built without considering the "Generosity" variable would have a better prediction.

Also looking at the linear regression coefficient indicate that "Family" and "Freedom" variables have a coefficient higher than 1. This might be as a result of some multicollinearity in variables or because no standardization has been done over the data.

Table 7.Regression model features

	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)
Coefficient	0.19531517	2.82123603	0.03475893	2.00103319	-0.50691269
Intercept	2.06248124				

5. Conclusion

In this study, I analyzed the relationship between variables such as Economy (GDP per Capita), Family, Health (Life Expectancy), Generosity, Freedom and Trust (Government Corruption) and a country's happiness and how they can assist in predicting a country's rank in world happiness analysis. I noticed that the generosity variable is not a good predictor of a country's happiness. I built a regression model to predict happiness scores. The model can be useful in predicting a country's score and rank in the coming years. By reviewing each variable's trend for a specific country such as Canada, we can predict the variable for the coming year e.x. 2021. Then by using the regression model we can estimate Canada's happiness score and future World rank.

It is suggested to future students that try building a model for predicting happiness scores after normalizing all data to see whether the prediction model could be improved. Moreover, it is suggested that an interrelationship between variables will be studied before starting analysis so that the variables with high correlation are omitted from the model. This might improve the predictive model of this study.