Project 1: World Happiness Report

Meenakshi Shankara
Bellevue University - Master of Science in Data Science
DSC680 Applied Data Science

06/24/2023

Table of Contents

INTRODUCTION

The World Happiness Report is a landmark survey of the state of global happiness. 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.

Business Problem

As the title says, the World Happiness report gives happiness scores and ranking for each of the countries. This provides a valuable insight into the least and most happy countries and what contributes to their rankings. This research is my attempt to dig a little more into what contributes to general happiness and do the factors on which the survey was conducted indeed are factors to one's happiness.

DATA

Datasets

The dataset that is being considered for this analysis is extracted from Kaggle website.

World Happiness Report | Kaggle

The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. 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 datasets from 2015-2017 also have a score called Dystopia Residual calculated from the happiness score of an imaginary country 'Dystopia' which has the world's least happy people. and thus sets a benchmark for the other country's scores and rankings. This score was removed from the datasets from 2018 onwards.

Data Dictionary

The following columns: GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption describe the extent to which these factors contribute to evaluating the happiness in each country.

Sno	Columns
1	Overall rank
2	Country or region
3	Score
4	GDP per capita
5	Social support
6	Healthy life expectancy
7	Freedom to make life choices
8	Generosity
9	Perceptions of corruption

The World Happiness Report is based on a structural model that includes variables such as happiness score, GDP, social support, life expectancy, freedom, generosity, and corruption. All variables are continuous variables and there is absence of missing values in the dataset.

Data Preparation

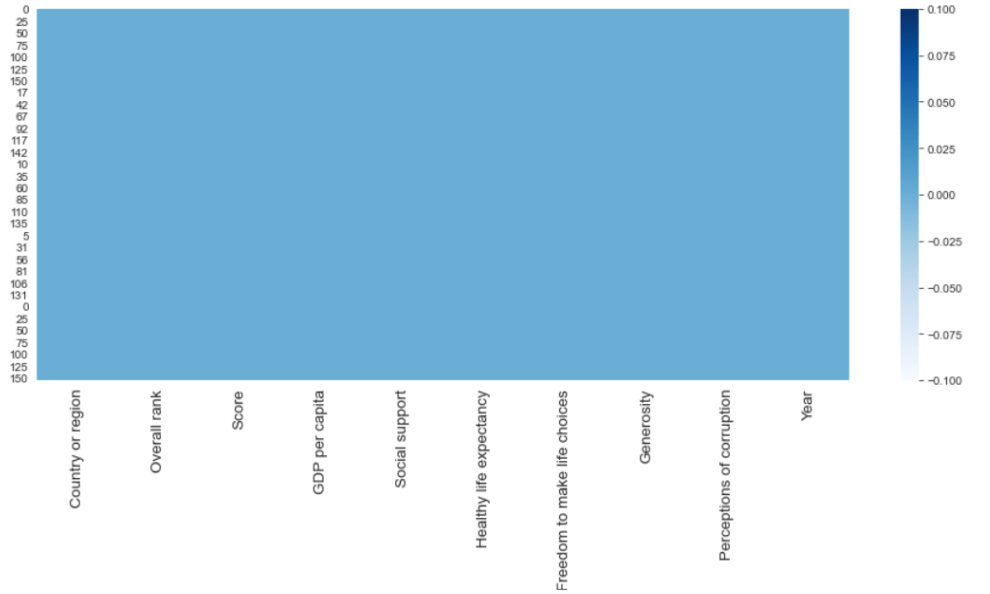
For research, all datasets from the years 2015-2019 will be merged. After merging the dataset has about 781 entries in total. This size is quite limited but can be utilized to provide a substantial prediction on factors that impact the happiness score.

The below transformations will be done on the dataset –

- Removed column Dystopia Residual, as it is not available in all datasets and is not a significant factor affecting the Happiness score.
- Rename columns for better appeal, understanding and to keep the names consistent across all years.
- Add the Year column to differentiate the datasets.
- Investigate the missing values in columns and remove them from the data frame.

```
# Double check to see if there are any missing values left
plt.figure(figsize = (16,6))
sns.heatmap(data = data.isna(), cmap = 'Blues')

plt.xticks(fontsize = 13.5);
```



Data Visualization

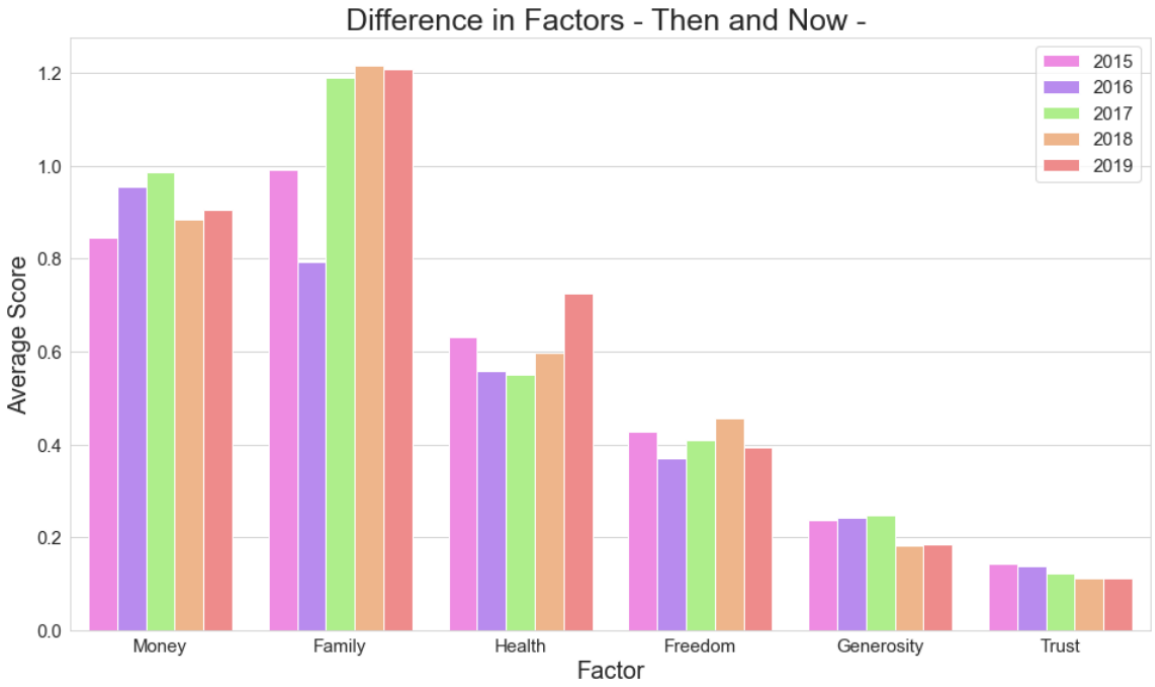
In the process of further analysis, we will now use the ‘describe()’ function in python to view the mean, min, max, percentile values grouping the data by year. This will help us in creating visualizations for comparing the statistical values across years.

```
data.groupby(by='Year')['Score'].describe()
```

Year	count	mean	std	min	25%	50%	75%	max
2015	158.0	5.375734	1.145010	2.839	4.5260	5.2325	6.24375	7.587
2016	157.0	5.382185	1.141674	2.905	4.4040	5.3140	6.26900	7.526
2017	155.0	5.354019	1.131230	2.693	4.5055	5.2790	6.10150	7.537
2018	155.0	5.366897	1.117433	2.905	4.4515	5.3580	6.15400	7.632
2019	156.0	5.407096	1.113120	2.853	4.5445	5.3795	6.18450	7.769

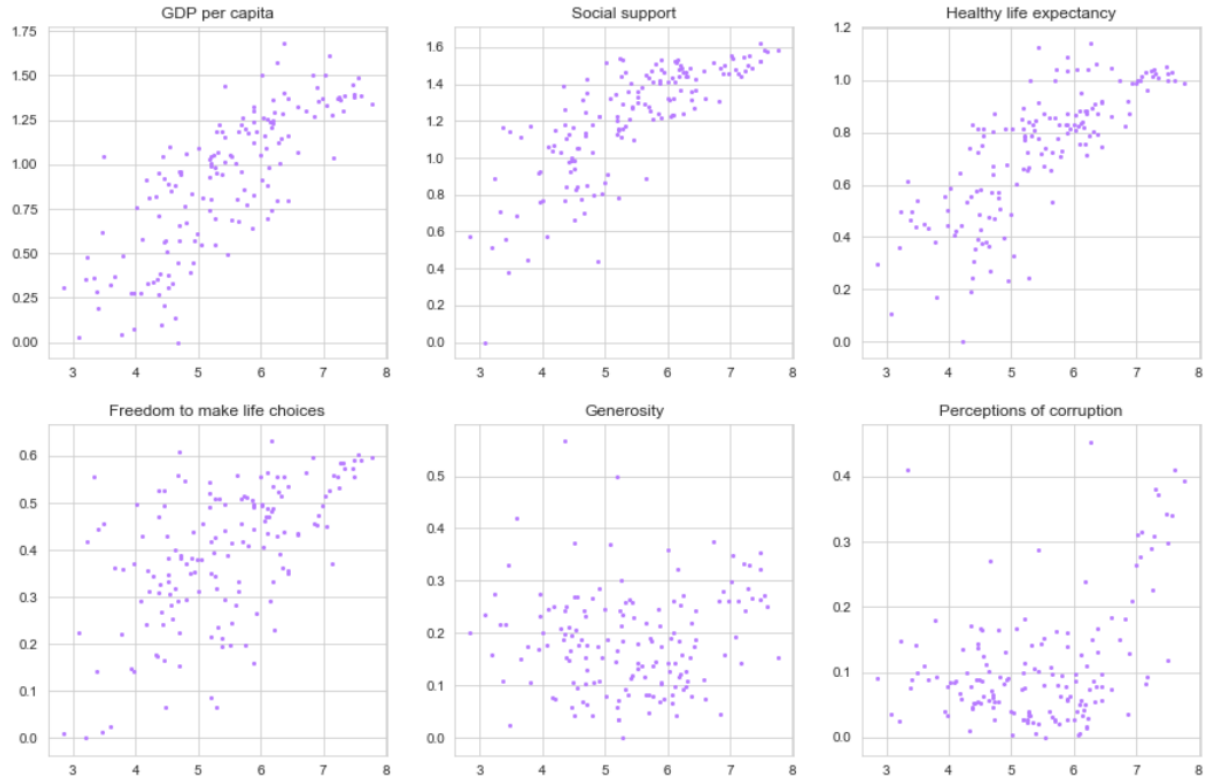
Next, expanding the same logic, the average value for each factor is calculated, factors being GDP per capita, Social Support, Healthy Life Expectancy, Freedom, Generosity, Perceptions of corruption. This will provide a view of which factor impacted the most for each year.

Barplots -



Overall, it seems like the year 2019 had more influence on the happiness score in factors like GDP and Family support. The year 2016 seems to have had lesser average scores in most factors.

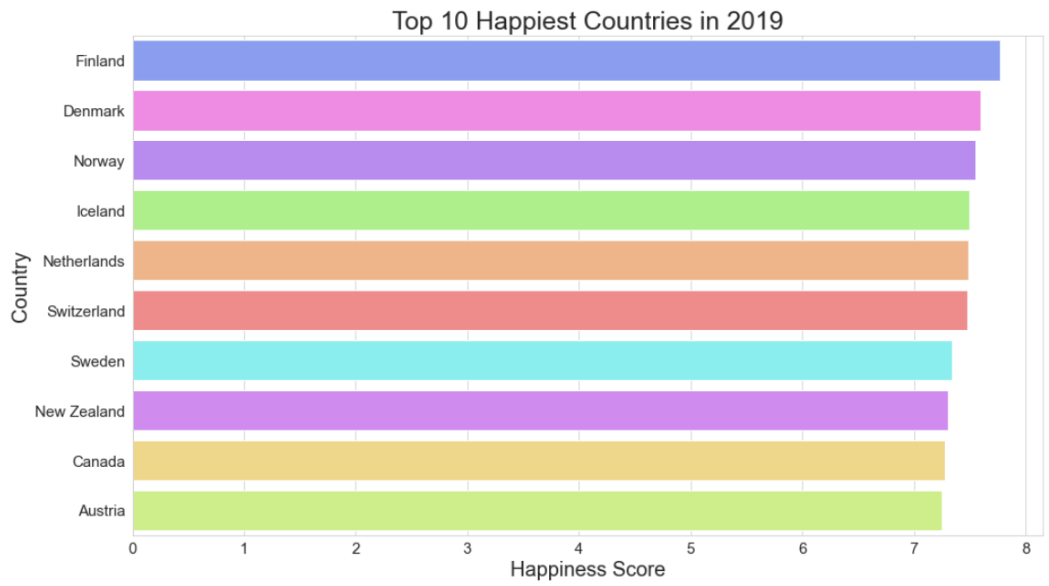
Next, plotting the value of each factor against the overall rank will provide a view of where the density of the values lay.



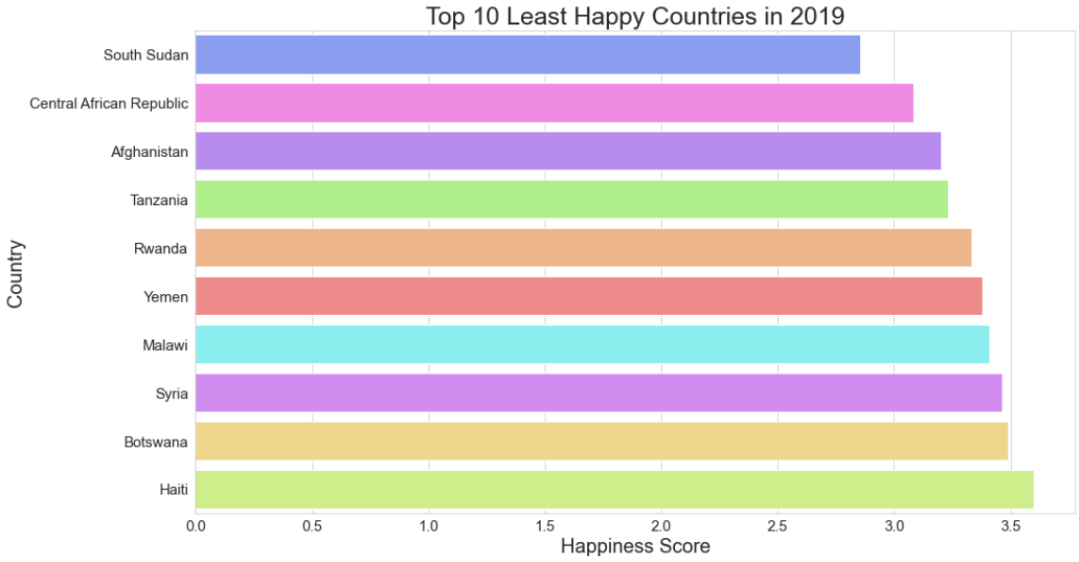
From the scatterplots we can see that GDP, Social Support, Healthy Life expectancy have a positive impact on the score.

To an extent Freedom also have a positive impact. Perception of corruption seems to have an impact. Maybe we need more data or better survey results to see a significant pattern. Generosity does not seem to impact the happiness score. Again, we may need more data for further analysis.

Based on the scores for each country, we can plot the Top 10 and Least 10 Happiest countries recorded in the year 2019.

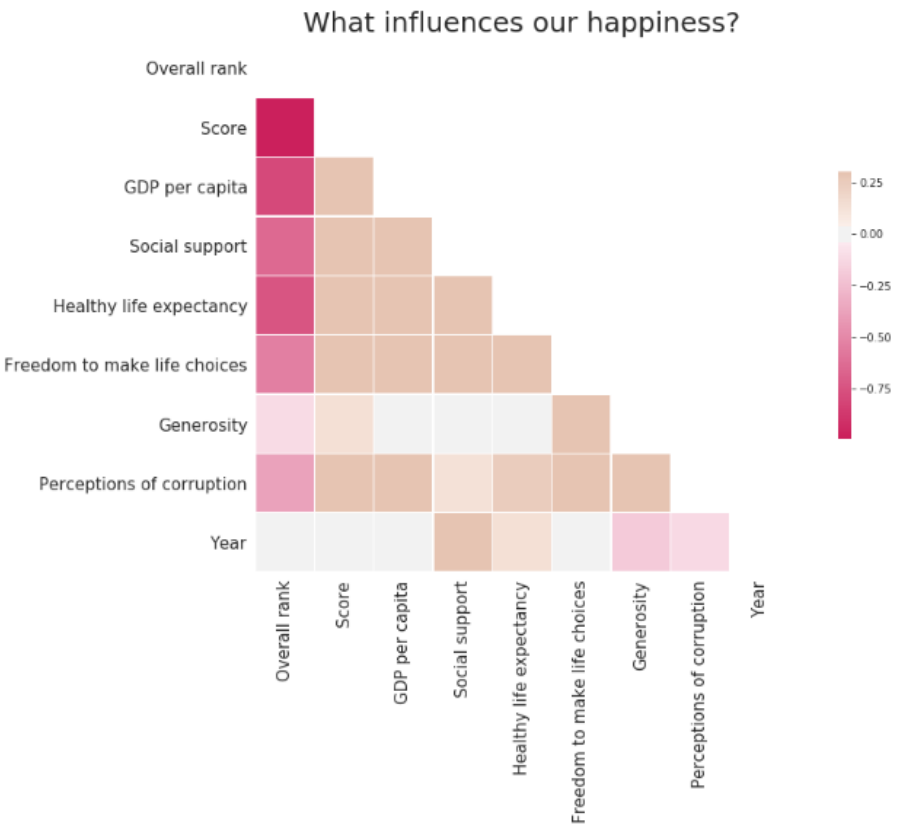


Finland, Denmark, and Norway lead the way on overall happiness scores.



It's no surprise that countries torn in wars, poor infrastructure, unstable governments have the least happiness scores.

Next, we calculate the correlation between each the happiness score with each factor and create a heat map on the correlation values.



We can see the Happiness score is most influenced by GDP and Good Health. There is also a positive correlation between Happiness, Freedom, and social support. Next is the perceptions of corruption. The happiness of a country seems to be least influenced by generosity among its citizens.

Modeling

The following modeling techniques are used to determine which modeling technique works best on this dataset and the features that are mostly related or correlated to the happiness score.

Linear Regression - computes the linear relationship between a dependent variable and one or more independent features.

Random Forest Regressor - A forest of randomly created decision trees, a combined output of individual decision trees to generate the final output.

Decision Tree - A series of sequential decisions made to reach a specific result.

Bayesian Linear - employs prior belief or knowledge about the data to “learn” more about it and create more accurate predictions.

From the Model score and Mean Average Error score, we see Bayesian Ridge has performed better than the other models with 83.24%. Next comes Linear Regression and Random Forest modeling methods with 82.98% and 82.54% respectively. The decision tree has not fared well; therefore, it may not be a good fit for this data analysis.

Model	Score	MAE
Linear Regression	82.98%	0.3769
Random Forest Regressor	82.8%	0.3658
Decision Tree	69.76%	0.5034
Bayesian Linear	83.24%	0.3723

Conclusion

What influences our general well-being?

- Financial well-being, in terms of GDP per capita
- Next is General health and long-life expectancy of the citizens. This translates to availability of medical care and attention.
- Freedom of expression and human rights is one of the top factors for happiness.
- The last one is generosity**.**

This report is helpful as it assesses the overall mood of a nation, as well as giving a glimpse of how it has evolved over time. If the GDP of a nation is not high or has not been showing consistent growth, then the country needs to focus on the same. This is useful in adjusting government policies, industries and prioritizing factors that influence the Happiness score.

This also gives an insight into what factors most influence the citizens of a country. Financial growth, good health and Freedom are of utmost importance for a nation to feel good.

Assumptions

The first assumption is that the dataset sample is significant enough to determine the happiness score of a country. According to the World Happiness report, about 1000 entries per country were considered for calculating the score.

The Second assumption is that the factors listed indeed impact the overall happiness/well-being of a country. While they do have significant weightage, happiness is a complex emotion that may depend on multitude of factors that have not been considered in the survey.

Ethical Assessment

One of the ethical considerations for this project is the consideration of results from the analysis in decision-making. Some of the conclusions made from this project's study could be incorrect or misrepresented due to insufficient or incorrect data. So, while sharing the outcome of this project to a larger audience, the underlying assumptions and data considerations should be shared.

1. No PII Data has been used for the analysis

2. All data sources are extracted from public domains, shared by organizations for educational purposes

3. All references are listed

References

World Happiness Report | Kaggle

A Structural Model of the World Happiness Report | R-bloggers

4 Takeaways From This Year's World Happiness Report (forbes.com)