A Study on World Happiness Index

BUDT 704-0507: Data Processing and Analysis in Python

Data Science Project By:
Group I
Ronak Shah
Hitarth Shah
Devni Shah
Vedant Kamat
Amoghvarsh Kulkarni
Sai Thanmayi Karpurapu

Mentor:
Dr Peng Huang

INTRODUCTION:

In an ever-evolving world, the pursuit of happiness is a fundamental aspiration for individuals and nations alike. The World Happiness Index, a comprehensive measure of well-being and life satisfaction, offers a unique lens through which we can analyze and understand the factors that contribute to happiness on a global scale.
Leveraging the power of Python, this project delves into the exploration and analysis of the World Happiness Index dataset, aiming to uncover patterns, trends, and insights that shape the happiness landscape across countries and over time.

Data Source:
https://worldhappiness.report/data/

The World Happiness Report is a landmark survey of the state of global happiness and we aim to study how happy we actually are.
For this study we have taken the data from the link displayed on screen. The time coverage for the dataset is from 2005 to 2023. There are approximately 165 countries which we have taken into consideration while doing our analysis.

Variables taken into Account while doing the study are as follows:

- **Country Name:** Name of Countries
- **Year:** Years taken into account
- **Life Ladder:** This is the indication of Happiness score. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you.
- **Log GDP Per capita:** The statistics of GDP per capita (variable name gdp) in purchasing power parity (PPP) at constant international dollar prices from World Development Indicators
- **Social Support:** This is to check whether there is anyone you can count on in times of help and other such variables that define the happiness of a person.
- **Healthy Life Expectancy at Birth:** It is the expected age of a child at birth.
- **Freedom to make Life Choices:** It is the national average of responses to the question: Are you satisfied or dissatisfied with your freedom to choose what you do with your life?
- **Generosity:** It is the national average of responses to the question: "Have you donated money to a charity in the past month?" on GDP per capita**.**
- **Perceptions of Corruption:** The measure is the national average of the survey responses to two questions: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?" The overall perception is just the average of the two 0-or-1 responses.
- **Positive Effect:** The average of three positive affect measures in GWP: laugh, enjoyment and doing interesting things in the Gallup World Poll waves 3-7.
- **Negative Effect:** The average of three negative affect measures in GWP. They are worry, sadness and anger, respectively.

MISSION OBJECTIVES:

- To identify what is the average happiness throughout all years in all countries
- To understand which are the Top happiest & Bottom least happy 10 countries
- To analyze the correlation between the variables that define the Happiness of a country
- To predict the Happiness of any particular country with the given variables by using Machine Learning

Data Cleaning:

1. We ensure data integrity by conducting a thorough examination for duplicates and affirming the absence of any duplicate entries in our dataset.
2. Removed data pertaining to the years 2005 and 2006, where the datasets exhibited less than 100 rows, optimizing dataset quality.
3. Excluded countries with less than 10 years' worth of data, resulting in the removal of entire corresponding rows.
4. Tackled missing data by utilizing the isna() method for each column, effectively identifying cells containing NaN values.
5. Employed a data-driven approach to handle missing values by imputing them with country-specific mean values, preserving the overall coherence and reliability of the data.
5. Post-imputation, we conducted a validation check, revealing that the 'country_healthy_life_expectancy' column for Hong Kong S.A.R. of China and Kosova, as well as the 'country_corruption' column for China and Turkmenistan, were entirely populated with NaN values.
6. Recognizing that the mean value would also be NaN, we made the informed decision to drop entire rows associated with these countries, ensuring data accuracy.
7. Achieved data consistency by eliminating both missing and duplicate values, resulting in a refined dataset ready for further analysis.

Data After Cleaning - <u>No missing values</u>

```
data.isna().sum()
```

```
Country name                      0
year                              0
Life Ladder                       0
Log GDP per capita                0
Social support                    0
Healthy life expectancy at birth  0
Freedom to make life choices      0
Generosity                        0
Perceptions of corruption         0
Positive affect                   0
Negative affect                   0
dtype: int64
```

<u>EXPLORATORY DATA ANALYSIS:</u>

Descriptive Statistics:

```
data.describe()
```

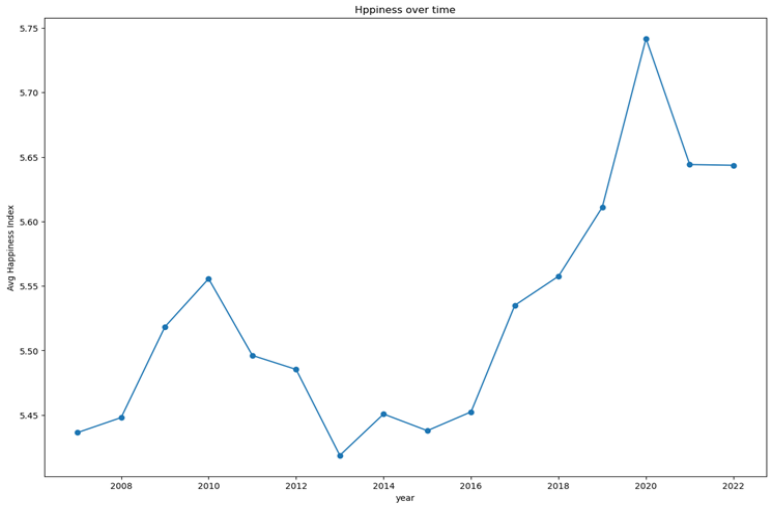
	year	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perceptions of corruption	Positive affect	Negative affect
count	1896.000000	1896.000000	1896.000000	1896.000000	1896.000000	1896.000000	1896.000000	1896.000000	1896.000000	1896.000000
mean	2014.663502	5.525224	9.442846	0.812467	63.836927	0.749804	-0.002600	0.739999	0.652260	0.273770
std	4.414495	1.132399	1.131834	0.119389	6.504014	0.138745	0.160735	0.188120	0.107706	0.085259
min	2007.000000	1.281000	5.527000	0.228000	17.360000	0.258000	-0.338000	0.035000	0.179000	0.083000
25%	2011.000000	4.669000	8.526000	0.747750	60.130000	0.659000	-0.117000	0.683000	0.568750	0.210000
50%	2015.000000	5.488500	9.548000	0.839000	65.325000	0.770000	-0.026000	0.797500	0.663000	0.265000
75%	2018.000000	6.375000	10.424250	0.906000	69.006250	0.861000	0.091000	0.867250	0.739250	0.326000
max	2022.000000	7.971000	11.664000	0.987000	74.475000	0.985000	0.703000	0.983000	0.884000	0.607000

In our exploratory data analysis (EDA), we began by visualizing the trend of the happiness index over the years. This provided insights into how overall happiness has evolved across different time periods.

Subsequently, we delved into the relationship between happiness and specific variables. We examined how perceptions, freedom to make life choices, and negative affect varied over time in relation to the happiness index. This approach allowed us to uncover potential patterns, correlations, or trends associated with these factors and their impact on overall happiness.

By conducting this EDA, we aimed to gain a comprehensive understanding of the dynamics between the happiness index and selected variables, shedding light on the nuanced factors that contribute to or influence happiness trends over the years.

1. **Happiness Index Trend Over the Years**



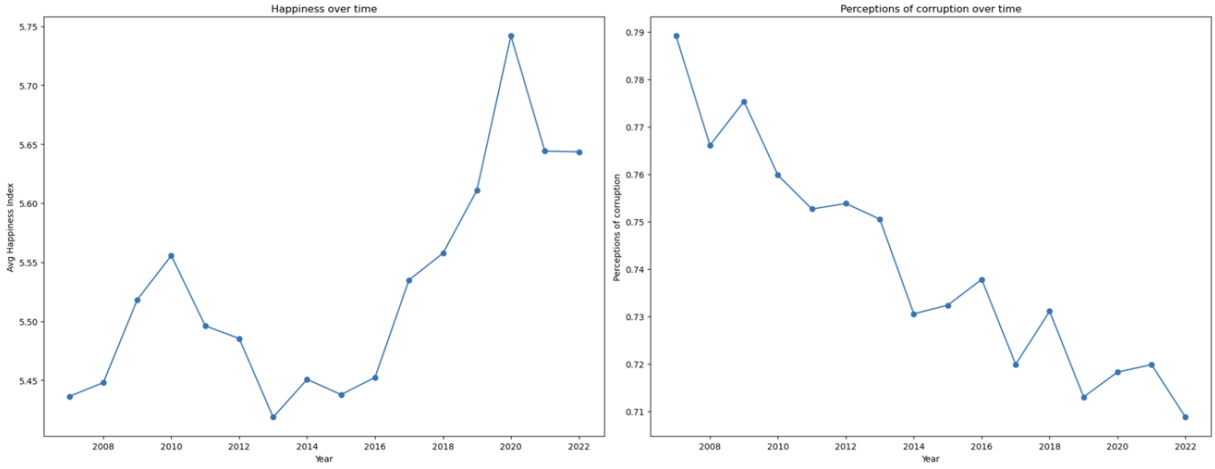
Insights drawn from the above plot reveal the following:

1. **Overall Positive Trend:** The happiness index demonstrates a general upward trajectory.
2. **Significant Increase Post-2008 Crisis:** Following the recovery from the 2008 financial crisis, there is a noticeable and substantial increase in the happiness index.
3. **Decline Amidst the Covid Era:** The happiness trend experiences a decline during the period coinciding with the Covid-19 pandemic.

Potential factors contributing to the drop in happiness from 2010 to 2014 include:

1. **Global Economic Conditions (Post-2008 Financial Crisis):** Numerous countries were still grappling with the aftermath of the 2008 financial crisis. Economic uncertainty, job losses, and a slow recovery could have collectively led to decreased happiness levels.
2. **Arab Spring (2010-2012):** The Arab Spring, characterized by protests and movements addressing political and economic concerns, resulted in significant social and political changes across the affected countries, potentially influencing happiness levels.
3. **European Debt Crisis (2010-2014):** Several European nations faced economic challenges and implemented austerity measures during the debt crisis. Elevated unemployment rates and economic difficulties may have adversely affected the well-being of citizens.

2. Happiness V/S Perception of Corruption Over the years

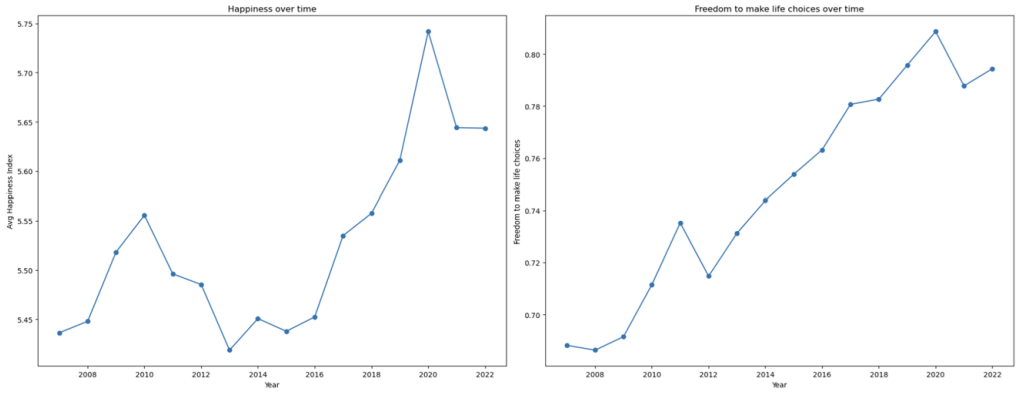


Observations derived from the depicted plot are as follows:

1. **Score Generation:** The score is derived from responses to questions assessing the prevalence of corruption, both within the government and businesses.
2. **Decreasing Trend:** There is a discernible downward trend over time in the corruption score. Interestingly, this aligns with the concurrent increase in the happiness index.

The inverse relationship between the corruption score and the happiness index suggests a potential correlation, implying that as perceptions of corruption diminish, overall happiness tends to rise.

3. Happiness V/S Freedom to make Life choices Over the years

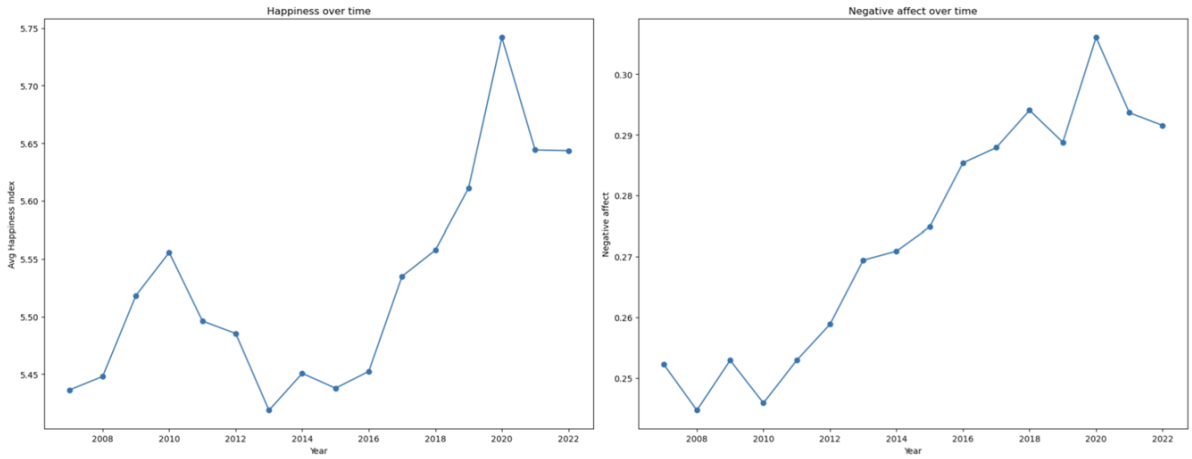


The observation from the provided plot indicates the following:

1. **Freedom of Choices Satisfaction:** The plot reflects individuals' satisfaction levels with the freedom of choices available to them.
2. **Increasing Trend:** Over time, there is a consistent and upward trend in the satisfaction levels related to the freedom of choices. Intriguingly, this positive trend aligns with the concurrent increase in the happiness index.

The correlation between the increasing satisfaction with freedom of choices and the rising happiness index suggests a potential relationship, implying that as individuals feel more satisfied with their freedom of choices, overall happiness tends to show an upward trajectory.

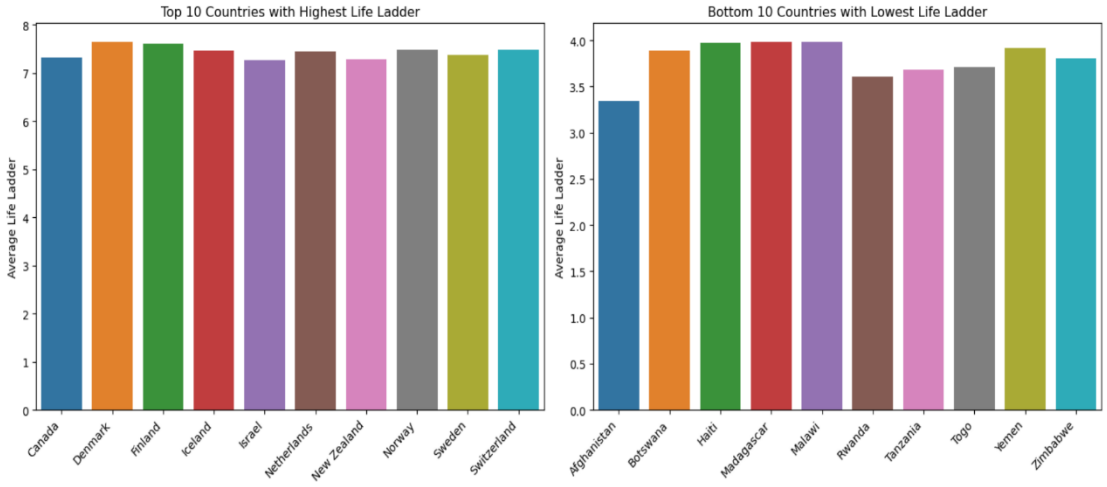
4. Happiness V/S Negative Affect Over the years



The data reveals an intriguing trend with potential implications for the happiness index. The negative effect, measured through responses to questions about worry, sadness, and anger, shows a surprising upward trend. Despite this increase in negative affect, the happiness index simultaneously demonstrates an overall upward trajectory.

This unexpected correlation prompts further exploration into the complex dynamics between negative affect and happiness. The coexistence of an increasing negative affect trend alongside a rising happiness index suggests that factors influencing happiness extend beyond the realm of negative emotional responses, indicating a nuanced relationship between the two.

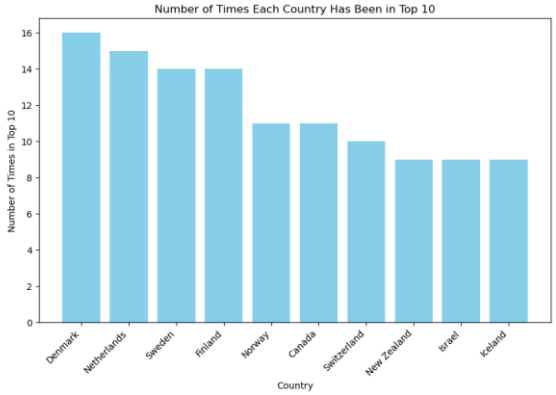
5. Bar Graph of top 10 and lowest 10 countries based on life ladder



The bar graphs depict the happiness levels of countries over the period from 2007 to 2023. Both graphs have a common X-axis listing the names of countries. On the right side, the graph focuses on the top 10 countries, while the left side features countries with the lowest happiness levels based on the life ladder,

represented on the Y-axis. The top 10 countries consistently exhibit significantly higher average happiness scores, hovering around 7.5, nearly double the average happiness scores of the lowest countries, ranging from 3.4 to 5.

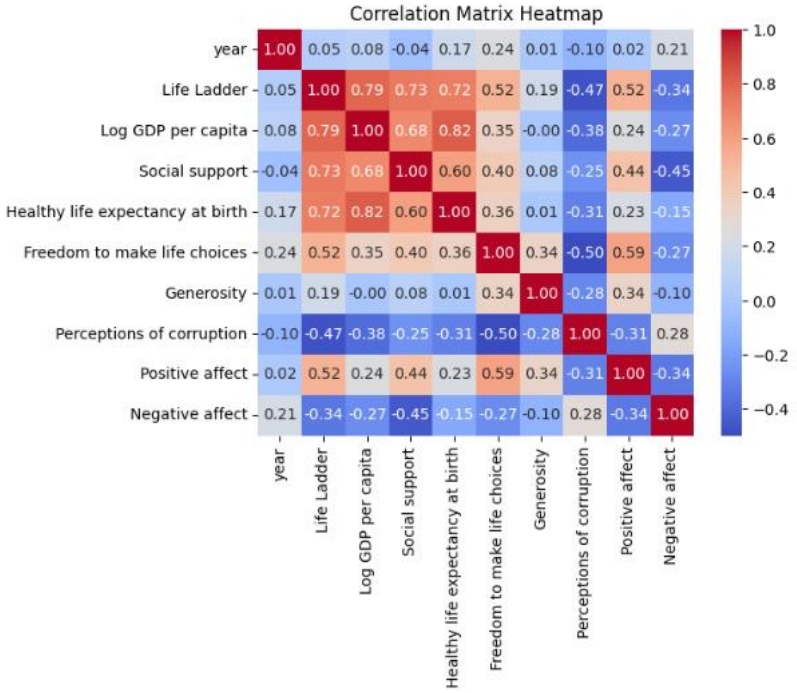
6. Bar graph of country vs Number of times in top 10



The bar graph in question displays countries on the X-axis and indicates the number of times each country has secured a position in the top 10 rankings over a span of 16 years. Notably, Denmark stands out as the most consistent performer, holding the 1st rank for all 16 years. This implies that Denmark has consistently maintained its top-tier position in terms of happiness throughout the entire period.

This observation aligns seamlessly with the accompanying graph depicting the happiness levels of the top 10 countries. In this graph, Denmark exhibits the highest happiness level of 7.5, surpassing all other countries. The correlation between Denmark's consistent top-ranking status and its elevated happiness level reinforces the idea that Denmark has not only maintained its position in the top 10 consistently but has consistently secured the top spot, reflecting a sustained high level of happiness across the analyzed years.

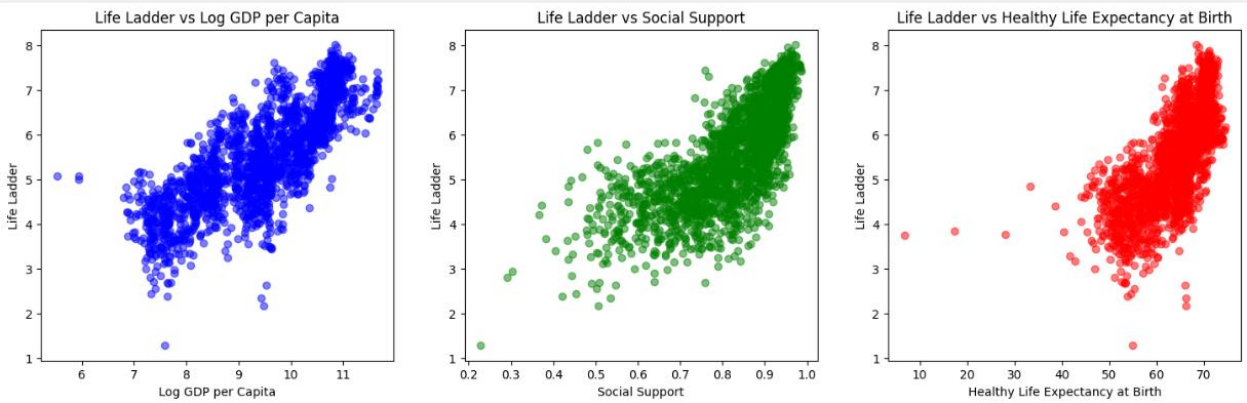
7. Correlation Matrix:



In examining the interplay of various socio-economic factors and their impact on well-being, our analysis via a correlation matrix heatmap has yielded insightful results. Notably, a strong positive relationship exists between citizens' life satisfaction (Life Ladder), economic prosperity (Log GDP per capita), and the support of social structures. Conversely, the perception of corruption is inversely related to these factors, underscoring its detrimental effect on societal well-being. The autonomy in life choices correlates positively with the prevalence of positive emotions, reaffirming the significance of freedom for psychological health. Interestingly, generosity appears to be independent of the other measured aspects, suggesting a multifaceted nature of altruistic behavior. The temporal aspect (year) revealed negligible direct correlations with the assessed factors, indicating a relatively stable interrelation across the time frame considered. These findings are pivotal for policy-

making, emphasizing the need for robust economic and social support systems while combating corruption to enhance national well-being.

8. Scatter Plot:



The scatter plots provided offer a visual representation of the association between the Life Ladder score and three key variables: Log GDP per capita, Social Support, and Healthy Life Expectancy at Birth. The Life Ladder score, a metric for overall life satisfaction, shows a positive trend with all three variables.

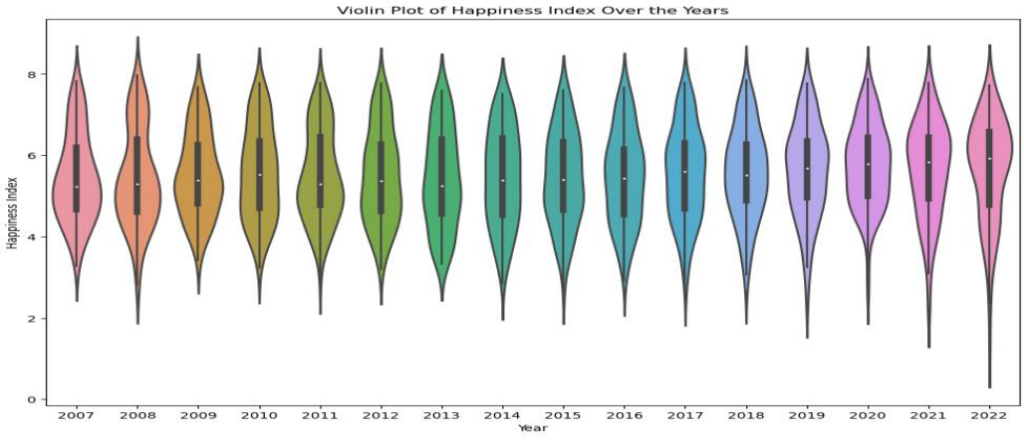
Log GDP per Capita: The first plot illustrates a positive correlation between economic output per person and life satisfaction, with higher GDP values aligning with higher Life Ladder scores.

Social Support: The second plot reinforces the importance of social relationships, indicating that individuals with more robust social support report greater life satisfaction.

Healthy Life Expectancy at Birth: The third plot highlights health as a significant factor, with longer healthy life expectancy associated with higher life satisfaction.

These plots underscore that economic, social, and health-related factors play a critical role in determining the quality of life and well-being of individuals. This data is invaluable for informing policy decisions aimed at enhancing the collective well-being of a population.

8. Violin Plot



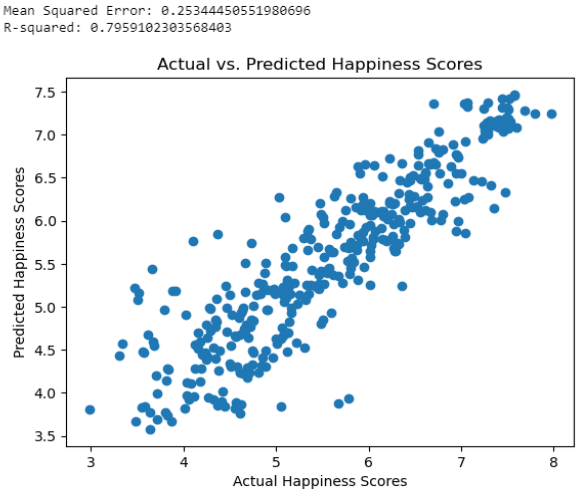
In the context of the Happiness Index over the years for different countries: The violin plot shows the distribution of Happiness Index values for each country across multiple years. Wider sections of the violin indicate years or values with higher density, while narrower sections indicate lower density. The box plot inside each violin provides additional insights into the central tendency and spread of the data. We can see a spread in the happiness indices of countries during major world events like the economic crisis of 2008. After 2019, due to COVID-19, Ukraine war etc., we can again see how spread out the violin plot is. It has resulted in happiness ranging very widely due to economic, political instabilities in the world.

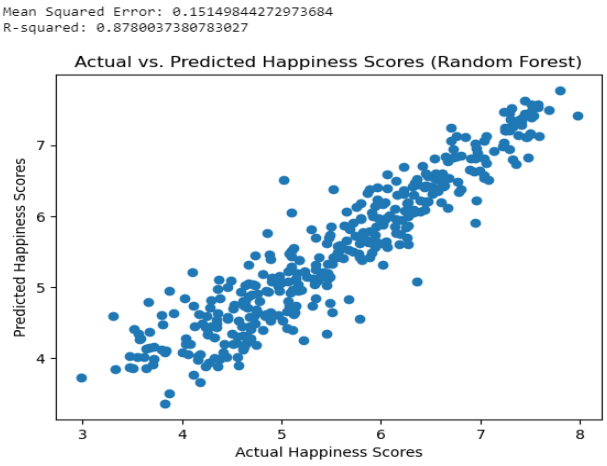
MACHINE LEARNING:

For performing model validation, we applied Machine Learning techniques on our World Happiness Index dataset.

The steps we followed for regression modeling were-

1. Defined features (X) and target variable (Y) i.e. Happiness score
2. Split the dataset into training and testing sets (80%-20% split)
3. Train the model using the training data
4. Made predictions on the testing set
5. Displayed model outcomes using scatter plot
6. Assessed model performance using metrics like squared mean error and R squared value





We evaluate our model by checking two important parameters- MSE (Mean Squared Error) and R-squared Score. A lower MSE value indicates better model performance. R-squared values close to 1.0 indicate a good fit, while values close to 0.0 indicate poor fit.

Both models are pretty good and accurate with high R-squared values of 0.795 for simple linear regression and 0.878 for random forest regression. They also have lower MSE values of 0.25 for simple regression and 0.15 for random forest regression.

Random forest has lower values of residuals, showing less differences between predicted and actual values. Random forest is an ensemble model where the building blocks are decision trees, which get trained on a random set of data, due to which it does not overfit and gives a much robust output as shown by its lower MSE value and higher R-squared value

How is this model useful to us? It can help in many ways such as-

1. **Predictive Analytics**: The trained model can be used for predictive analytics to estimate future happiness scores based on changing conditions.
2. **Understanding Feature Importance**: The feature importance scores provide insights into which factors contribute the most to happiness scores.
3. **Monitoring Changes Over Time**: The model can be applied to new data to monitor changes in happiness trends over time.
4. **Targeted Interventions**: With insights into feature importance, interventions can be targeted at improving specific aspects that have a significant impact on happiness. This could involve targeted social programs, health initiatives, or measures to enhance personal freedoms and choices.
5. **Monitoring Changes Over Time**: The model can be applied to new data to monitor changes in happiness trends over time. Tracking shifts in feature importance and happiness scores provides a dynamic view that can inform ongoing strategies and interventions.
6. **Resource Allocation**: Organizations and governments can allocate resources more efficiently by understanding which factors are most influential in determining happiness. This can lead to more effective resource distribution for the well-being of the population.
7. **Predictive Analytics**: The trained model can be used for predictive analytics to estimate future happiness scores based on changing conditions.

CONCLUSION:

The analysis of global happiness metrics reveals a heartening trend; the average happiness score for the leading ten countries stands at an impressive 7.5. Additionally, the general trajectory of happiness over the years suggests a positive incline, hinting at improving global sentiments. Core determinants such as economic output, social frameworks, and health parameters emerge as pivotal influencers in shaping a nation's happiness quotient. Notably, regional disparities in these happiness trends are apparent, reflecting the unique socio-economic and political landscapes that characterize different geographies. The utilization of regression models has proven to be particularly effective, offering a significant degree of accuracy in predicting national happiness levels. This not only validates existing well-being metrics but also encourages deeper exploration into the science of happiness and its determinants.

RECOMMENDATIONS:

In light of the findings, it is recommended that health care be made a universal mandate, assuring all citizens equitable access to essential medical services, which is a cornerstone for fostering national happiness. Economic strategies should also be revised to prioritize growth in GDP per capita, focusing on bolstering investment and developing infrastructure as means to this end. Moreover, to sustain and improve the happiness index further, it is imperative for policies to nurture work-life balance. This could involve instituting flexible working arrangements and comprehensive parental leave policies, thereby addressing the multidimensional nature of happiness and contributing to the well-being of the populace. By implementing these recommendations, governments can make significant strides towards enhancing the collective contentment and prosperity of their nations.