Creating Data Visualizations

Individual Contribution			
CWID	Name	Contribution(description)	Percent Contribution
A20563450	Longbo Xiang	Write code and report section 4,2	50%
A20563459	Jun Yang	Write report sections 1, 2 and 3	50%

1. Dataset Description

1.1 Data Source and Background

The dataset used in this report is derived from the **2024 World Happiness Report** and is stored in CSV format. It records the happiness levels and influencing factors of **143 countries** worldwide. The dataset is structured as a **record-type data matrix**, where each row represents a country or region, and each column represents a variable measuring happiness or its related factors.

1.2 Key Variables

The dataset encompasses multiple dimensions, including economic, social, and health factors. The key variables include but are not limited to:

Country Name: Name of the country.

Happiness Rank: Rank of the country

Happiness Score: Score of the country.

Upperwhisker : Upper score.

Lowerwhisker : Lower score.

Economic Indicator (GDP per capita): GDP.

Social Support: Score from social support.

Healthy Life Expectancy: Score from Life Expectancy.

Freedom to Make Life Choices: Score from Freedom.

Generosity: Score from Generosity.

Perceived Corruption: Score from Perceptions of corruption.

1.3 Attribute Types

Nominal Attribute: Country Name.

Ordinal Attribute: Happiness Rank.

Ratio Attribute: Upperwhisker, Lowerwhisker.

Continuous Attributes: Happiness Score, Economic Indicator (GDP per capita), Social Support, Healthy Life Expectancy, Freedom to Make Life Choices, Generosity, Perceived Corruption.

1.4 Data Types

Most variables in the dataset are **continuous numerical data**, while some categorical information (such as Country Name) is **discrete data**. The numerical differences between variables allow for the discovery of potential relationships through statistical analysis and visualization techniques.

1.5 Data Preprocessing Requirements

During the subsequent data analysis, **preprocessing** is required to handle missing values, outliers, and data distribution issues. This ensures data completeness and accuracy, preventing biases in analysis results due to poor data quality.

2. Visual Presentation and Method Explanation

For describing happiness indicators and exploring their influencing factors, we choose appropriate graphs to describe them accordingly.

Overview of specific methods: Firstly, corresponding **data preprocessing** is carried out based on data characteristics, such as aggregation,sampling, standardization and outlier removal, etc. Then, the library is imported and corresponding functions or machine learning models (random forest) are called for visual analysis.

2.1 World Happiness Index Map

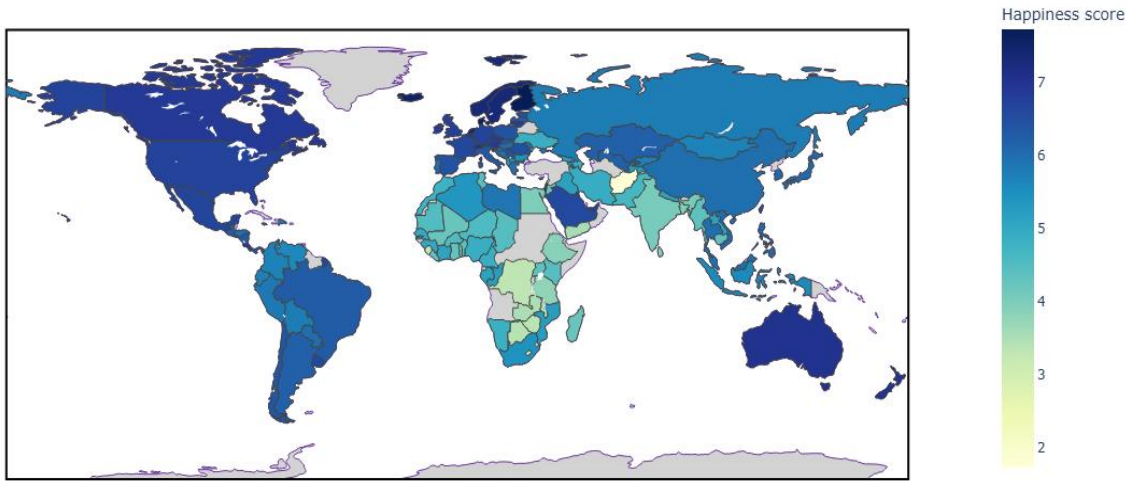


Figure 1: World Happiness Index Map

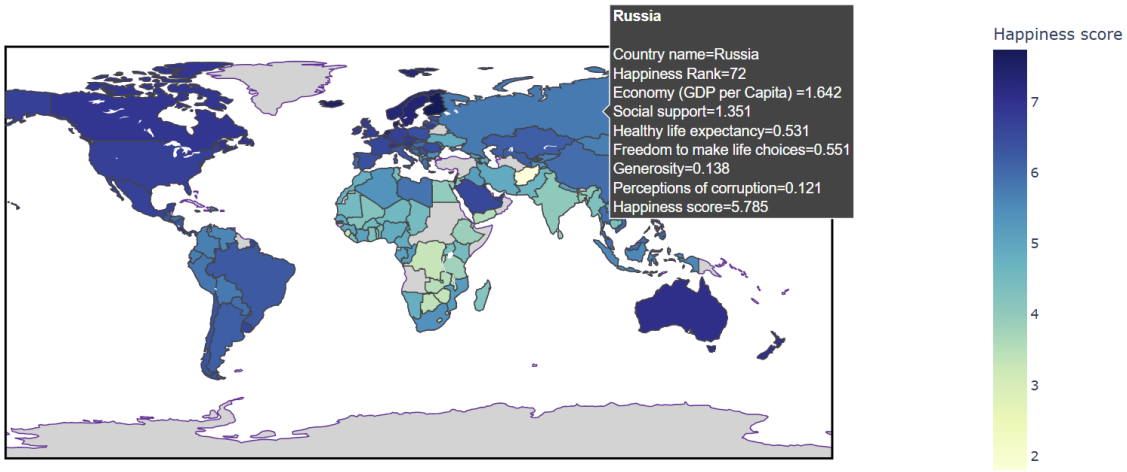


Figure 2: Interactive Map

Figure 1 visually displays the happiness index levels of different countries around the world through color depth. The color bar on the right corresponds to different happiness index values, with darker colors indicating higher happiness index values and lighter colors indicating lower happiness index values. Countries with high happiness indices are concentrated in western Europe, Oceania, and other regions, where countries typically perform well in terms of economy, social welfare, and other aspects; However, the happiness index of some countries in Africa and South Asia is relatively low.

By moving the mouse to that country, we can see the corresponding information for that country, which is shown by Figure 2.

Method Explanation

Firstly, for missing values, **use 0 to fill in**, and other graphs are also based on this operation.We import **plotly. express** (px) to create an interactive visual chart, and then use the **px. choropleth function** to create a color map based on geographic regions. Among them, the locationmode is set to 'country names' to match by country name, color specifies the column used for color mapping (happiness index score), color_continuous_scale selects the color gradient scheme, hover_name sets the content displayed when hovering the mouse, and hover_data defines the additional indicator data displayed when hovering.

Then, using the **fig.update_geos** method, we set the display style and color of coastline, land, and ocean, select the projection type, and configure the display properties of longitude and latitude axes, such as whether to display gridlines, scale

intervals, and ranges. We use the **fig.update_layout** method to set the geographic coordinate system framework, coordinate axes (including title, font, line style), chart size, title, and background color in detail to achieve a beautiful and easy to read effect.

2.2 Distribution of Average Happiness Score

Figure 3: Distribution of Average Happiness Score

This histogram shows the distribution of happiness indices in various countries around the world. On the central tendency, around 6 is the mode range, where the happiness index of most countries is concentrated. The distribution pattern is similar to a unimodal bell shaped curve, approaching normal but asymmetric. The curve in the low happiness index area is flat and the distribution of countries is scattered, while the curve in the high happiness index area decreases rapidly and there are few countries. The range of values shows that there are significant and substantial differences in happiness indices among countries. From the changes in width and height of the graph, it can be seen that the low happiness index interval has a high degree of dispersion, the high happiness index interval has a low degree of dispersion, and the moderate interval has a low concentration of dispersion. Overall, there are significant differences in happiness levels among countries around the world.

Method Explanation

Firstly, import the pandas library for data processing, matplotlib. pyplot for basic plotting functions, and seaborn for creating more visually appealing statistical charts.

Then, **sns.histplot** (data=df, x='Happiness score', bins=20, kde=True) uses the 'Happiness score' column in the DataFrame (df) as data, draws a histogram, divides the data into 20 intervals (bins=20), and simultaneously draws a kernel density estimation curve **(kde=True)** to display the distribution pattern of the data.

2.3 Rank **Horizontal** Bar Chart

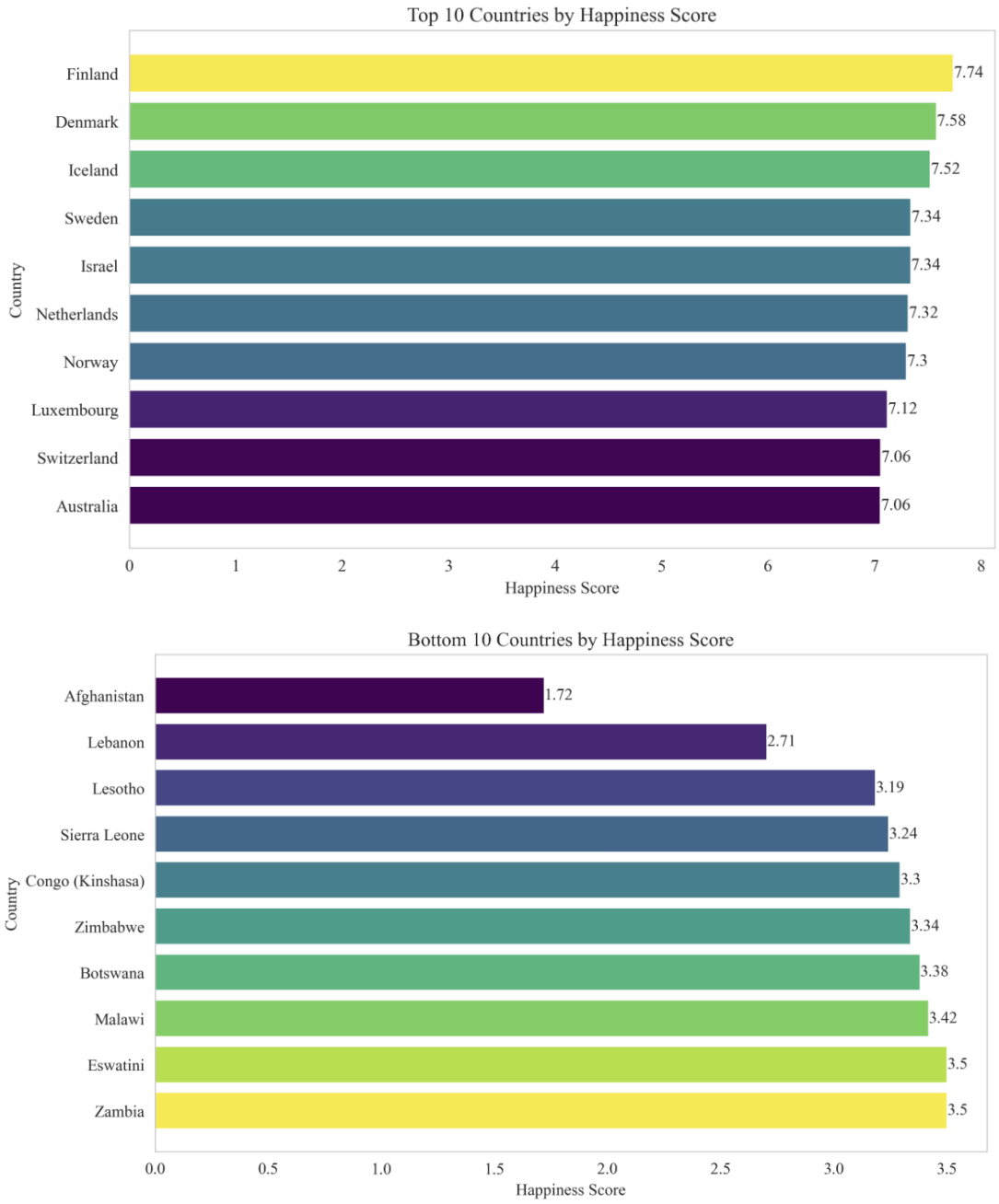


Figure 4: Rank by Happiness Score

Figure 4 presents two horizontal bar charts displaying the top 10 and bottom 10 countries in terms of Happiness Rank. The bars are arranged in descending order for the top-ranking countries and ascending order for the lowest-ranking countries. The

length of each bar represents the Happiness Score of the corresponding country, with the exact score labeled on the right. A continuous "viridis" colormap is applied, where the Happiness Score is normalized to determine the color intensity: higher rankings appear brighter, while lower rankings appear darker. Finland ranks highest with a Happiness Score of 7.74, whereas Afghanistan ranks lowest with a score of 1.72.

Method Explanation

To extract the top and bottom 10 countries based on happiness ranking, we use **df.nsmallest**(10, 'Happiness Rank') and **df.nlargest**(10, 'Happiness Rank'), respectively. These subsets are then sorted using **sort_values**(by='Happiness Rank', **ascending=False**) and sort_values(by='Happiness Rank') to ensure that in the horizontal bar chart, the highest-ranking countries appear at the top.

Then the Happiness Scores are normalized using **Normalize()** based on their range (minimum to maximum values). The normalized values are then mapped to colors using **cm.get_cmap('viridis')** and **cm.get_cmap('viridis', 10).reversed()**, allowing each bar's color to dynamically reflect its score.And we use **plt.barh** to create horizontal bar charts, effectively visualizing the ranking of countries by happiness score.

2.4 Combination Diagram

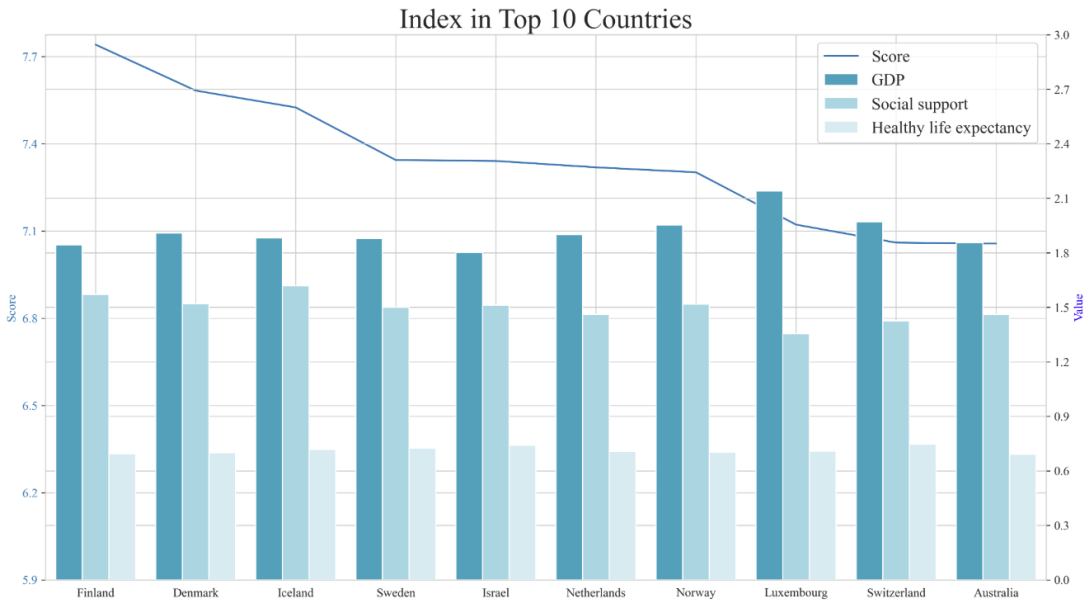


Figure 5: Combination Diagram

By combining **line charts and grouped bar charts**, display the top ten countries in terms of happiness index GDP The situation regarding social support and healthy life expectancy.The happiness index is represented by a blue line. GDP is represented by dark blue bars. The values are relatively concentrated, with most countries having a

GDP between 1.8-2.1, with Luxembourg having the highest GDP among these countries. Social support is represented by light blue bars. The numerical distribution is also relatively concentrated, with relatively small differences among countries, fluctuating around 1.5. Healthy life expectancy is represented by a light blue bar. The values are generally lower than other indicators, ranging from 0.6 to 0.9, and the gap between countries is not significant.

Method Explanation

Use the melt method to reshape the top_10 data box, including 'GDP' and ' Perform long format conversion on the data in columns' Social support 'and' Healthy life expectation '.
Set bar chart color: Define a color list **bar_comors** containing a light blue gradient, used to set colors for different columns in grouped bar charts to enhance visualization.
Use the **barplot** function of the Seaborn library to draw a grouped bar chart on the other axis ax2 of the dual axis. The x-axis represents the country name, the y-axis represents the value of the corresponding indicator, and the hue parameter is grouped according to different indicators ('Metric '), with the color set to bar_comors.

2.5 Correlation Heat Map

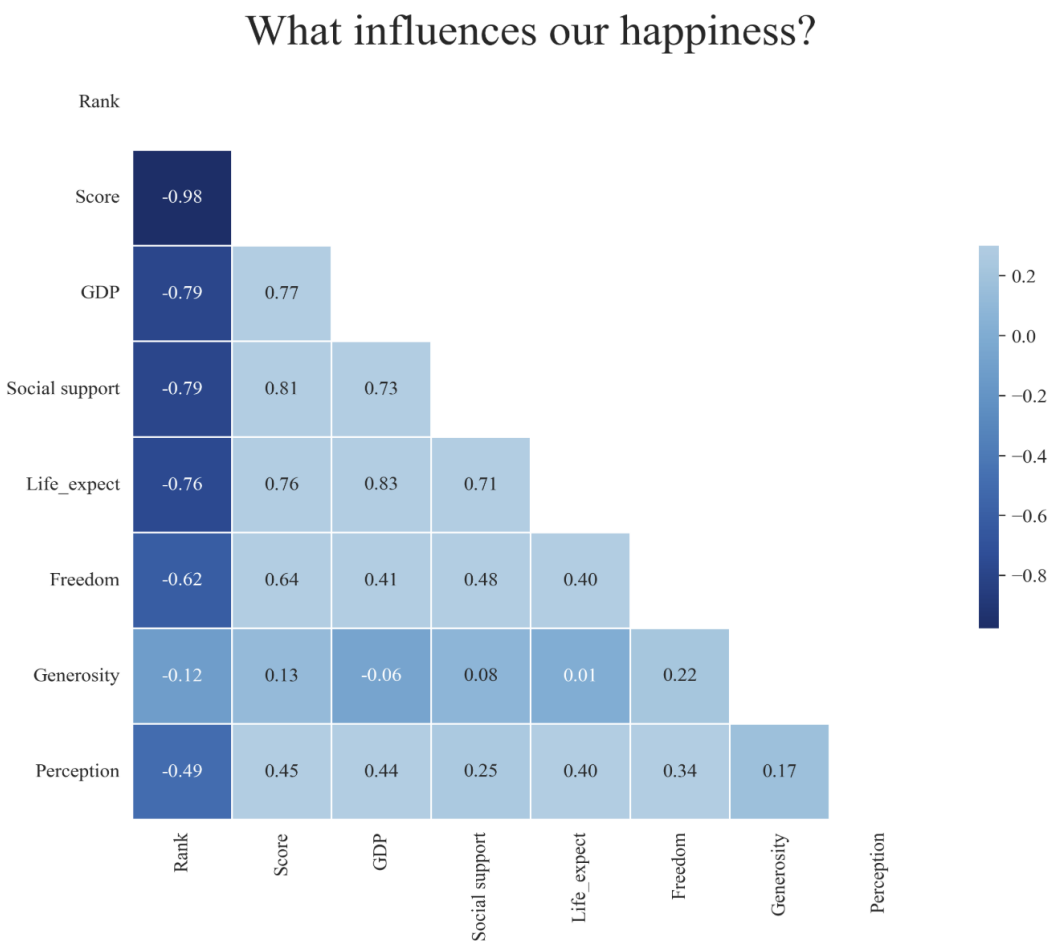


Figure 6: Correlation Heat Map

Figure 6 presents a heatmap illustrating the pairwise Pearson correlation coefficients among eight indicators: Happiness Rank, Happiness Score, GDP, Social Support, Healthy Life Expectancy, Freedom to Make Life Choices, Generosity, and Perceptions of Corruption. Use the following formula:

$$\mathrm{Cov}(X,Y) = E\{[X - E(X)] \cdot [Y - E(Y)]\}$$

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X) \cdot D(Y)}}$$

where X and Y represent random variables, $E(X)$ and $E(Y)$ denote their expected values, $D(X)$ and $D(Y)$ are their variances, $Cov(X,Y)$ is their covariance, and ρ_{XY} is the Pearson correlation coefficient.

Since the correlation matrix is symmetric about the main diagonal, only the lower triangular part is displayed in the heatmap. Each cell in the heatmap is labeled with the corresponding correlation coefficient, while the color intensity represents the strength of the correlation.We observe a strong positive correlation between Happiness Score and Social Support and between GDP and Healthy Life Expectancy.

Method Explanation

Firstly,compute the Pearson correlation coefficients and generate a symmetric correlation matrix. A masking array is created using **np.triu** to hide the upper triangular part, ensuring that only the lower triangular section is displayed. Finally, we use **sns.heatmap** to visualize the correlation matrix as a heatmap.

2.6 Feature Importances

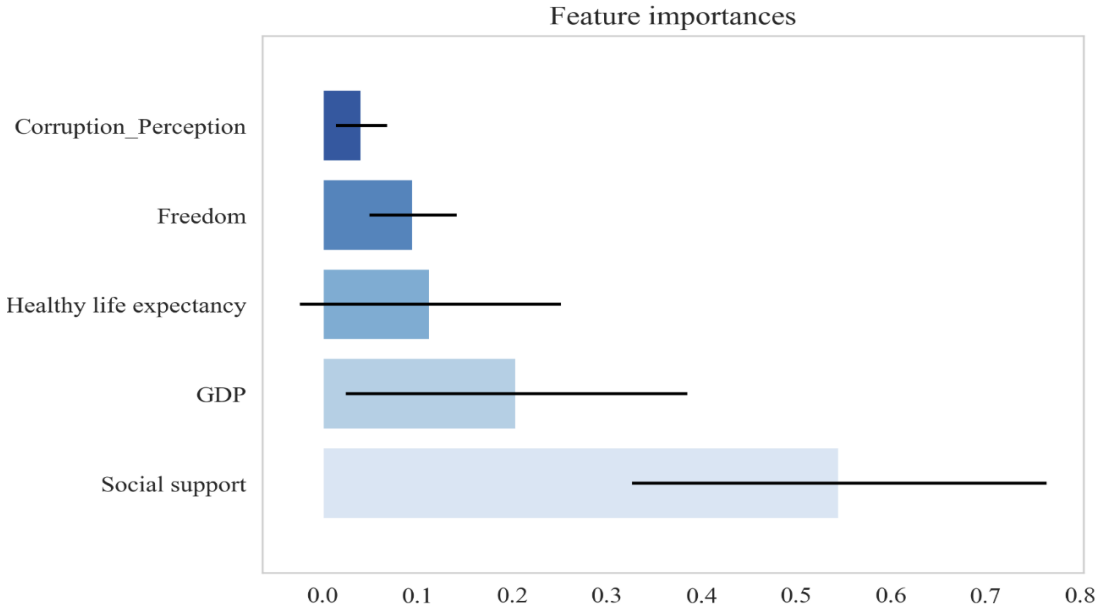


Figure 7: Feature Imporrtances

Figure 7 presents the feature importance ranking in the Random Forest model, with error bars indicating the variability of importance scores across different decision trees. In the horizontal bar chart, the length and color intensity of each bar represent the feature's importance score. Features are ranked in ascending order of importance, with lighter colors indicating higher importance scores. The error bars (xerr) attached to each bar show the standard deviation of the importance scores across different decision trees. The features ranked from highest to lowest importance are: Social Support, GDP, Healthy Life Expectancy, Freedom to Make Life Choices, and Perceptions of Corruption.

Method Explanation

First, five features related to Happiness Score prediction—GDP, Social Support, Healthy Life Expectancy, Freedom, and Perceptions of Corruption—are selected from the dataset. These features are then standardized using **StandardScaler**.

Next, the feature importance scores are obtained from the **Random Forest model** (feature_importances_), which measures the average contribution of each feature to reducing impurity across all decision trees. The standard deviation of each feature's importance score is computed using **np.std()**.

Finally, a horizontal bar chart is plotted using **plt.barh()**, incorporating **error bars** (xerr=std[indices[i]]) to visualize the variability of feature importance scores across different trees.

2.7 SHAP Summary Plot

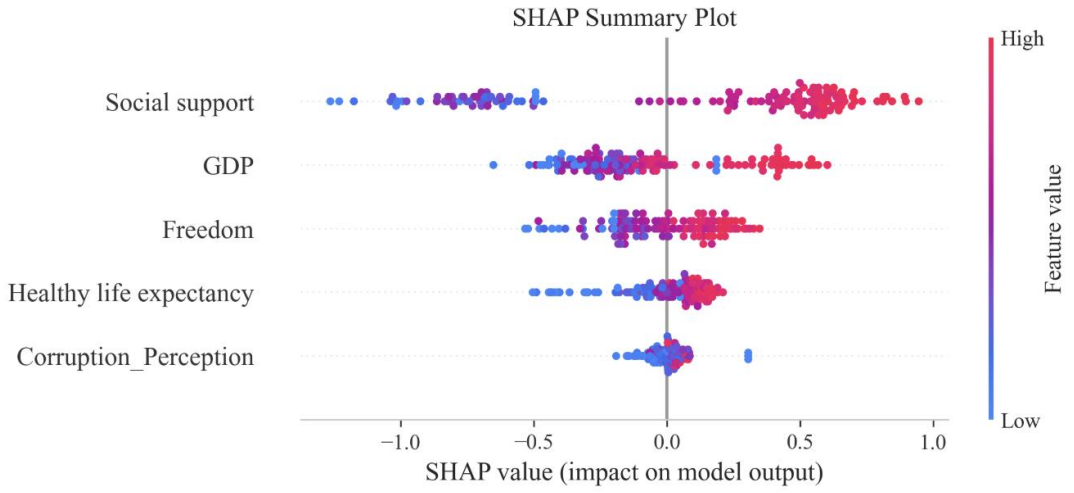


Figure 8: SHAP Summary Plot

Figure 8 presents the SHAP Summary Plot, illustrating the contribution and

directional impact of each feature on the Random Forest model's prediction of Happiness Score. The SHAP is explained as follows:

SHAP (SHapley Additive exPlanations) is a model-agnostic method based on game theory. It calculates the marginal contribution of each feature to the model output, providing explanations both at a global and local level for "black-box" models. SHAP treats all features as "contributors" and computes their marginal contributions using Shapley values, forming an additive explanatory model. The Shapley value is calculated as follows:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\} - v(S))$$

where N represents the full set of features, S is any subset of features excluding feature i, $|S|$ is the number of features in subset, $v(S)$ represents the model output for subset S, $v(S \cup \{i\})$ represents the model output when feature i.

From the plot, it is evident that Social Support has the widest color spread and contains more red points, indicating that it has the most significant positive impact on Happiness Score. This suggests that in 2024, social support is the primary factor influencing happiness, followed by economic factors such as GDP. The results highlight that good governance is essential for people's well-being, and all these factors play a crucial role in achieving effective governance.

Method Explanation

We use **RandomForestRegressor** to build a **Random Forest regression model** with **100 decision trees** (n_estimators=100) and a **fixed random seed** (random_state=42). The standardized feature set X and target variable y are fed into the model for training. To interpret the trained model, we use **shap.TreeExplainer** to compute the SHAP values for each feature in each sample, reflecting their local contributions to the model's predictions.

Finally, we generate the SHAP summary plot using **shap.summary_plot**, where **color and point distribution** reveal the direction and strength of each feature's impact on Happiness Score predictions.

2.8 Multivariate Scatter Plot Matrix

Figure 9: Multivariate Scatter Plot Matrix

Figure 9 presents a multivariate scatter plot matrix, illustrating the pairwise relationships among four key indicators: Happiness Score, GDP, Social Support, and Healthy Life Expectancy. The diagonal elements of the matrix display kernel density plots, showing the distribution of happiness levels for each indicator. The off-diagonal elements contain scatter plots that visualize the relationships between pairs of variables, with color representing different happiness categories: High Happiness (blue), Medium Happiness (orange), and Low Happiness (green). From the plot, we observe that the Medium Happiness category exhibits the most concentrated distribution across all four indicators.

Method Explanation

First, we define the **categorize_score** function to **aggregate happiness scores** into three categories:
High Happiness (Score ≥ 7, **Blue**)
Medium Happiness (Score between 5 and 7, **Orange**)
Low Happiness (Score < 5, **Green**)
A **custom color dictionary (custom_palette)** is then created to assign specific colors to these three happiness categories.

Finally, we use **sns.pairplot** to generate the multivariate scatter plot matrix, applying the **hue='Happiness Category'** parameter to color the data points according to their happiness category, enabling a clear visualization of the relationships between the selected indicators.

2.9 Box Plot of Happiness - Related Matrics

Figure 10: Box Plot of Happiness - Related Matrics

Figure 10 presents a box plot visualization of the distribution of six happiness-related indicators: GDP, Social Support, Healthy Life Expectancy, Freedom, Perceptions of Corruption, and Generosity. Each box plot corresponds to one indicator, displaying its median, interquartile range (IQR), and outliers.
From the figure, we observe that GDP and Social Support have relatively high overall scores and exhibit a more dispersed distribution, whereas Perceptions of Corruption and Generosity have lower overall scores and a more concentrated distribution. Additionally, the medians for all indicators are roughly centered within the boxes, indicating that their distributions are fairly symmetric.

Method Explanation

We conducted **feature construction.**We convert the wide-format dataset into a long-format structure, transforming the selected indicators into two columns:

"Metrics" (indicator names), "Values" (corresponding scores).

Finally, we use Seaborn's **sns.boxplot** function, setting x='Metrics' and y='Values' to specify the indicator names and their values, creating the box plots to visualize the data distribution.

2.10 Multi-Indicator Trend Chart

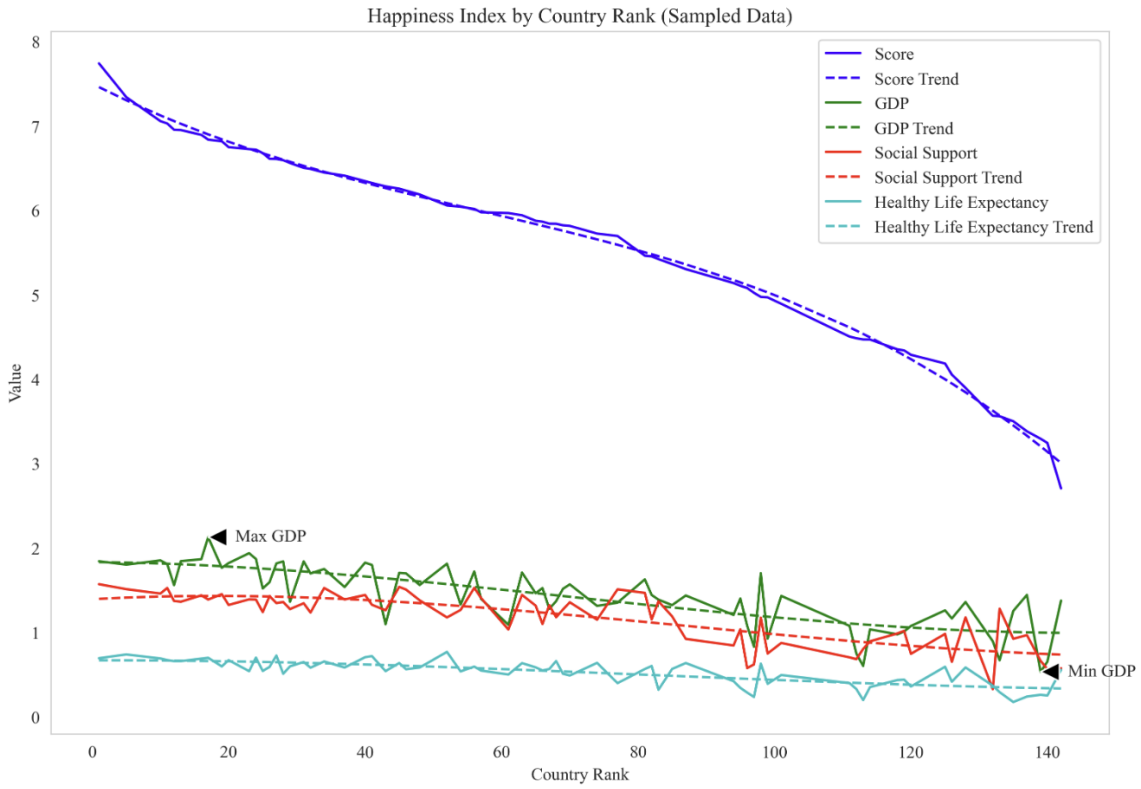


Figure 11: Multi-Indicator Trend Chart

Figure 11 presents trend curves illustrating how different indicators change with country rankings. The plot shows the variation of Happiness Score, GDP, Social Support, and Healthy Life Expectancy across rankings, with minimum and maximum GDP values explicitly marked along the GDP curve.
This visualization provides insights into the relationship between country rankings and key indicators. From the figure, we observe that the Happiness Score curve is steep at both high and low rankings but relatively flat in the middle range. In contrast, GDP, Social Support, and Healthy Life Expectancy do not exhibit a clear correlation with rankings and show significant fluctuations, especially among lower-ranked countries.

Method Explanation

Firstly, perform data preprocessing
Random sampling: Randomly select 50% of the original data df using the sample

method

Outlier removal: Use the interquartile range (IQR) method to identify and remove outliers in the GDP, Score, Social Support, and Healthy Life Expectancy columns of the sample.

Handling missing values: Use the dropna method to remove rows in the sample that contain missing values.

Grouping aggregation: Group by Rank and calculate the mean of each numerical column within each group.

Draw the original line chart to show the changes in indicators with country rankings. Perform third-order polynomial fitting, calculate the fitting curve using the np. polyfit and np. poly1d functions, and plot it as a dashed line to represent the overall trend of the indicator.

2.11 Scatter Diagram

Figure 12: Scatter Diagramt

Figure 12 illustrates the relationship between GDP per capita and Happiness Score across countries. Each point represents a country, with the x-axis denoting GDP per capita and the y-axis representing the Happiness Score.

Data points are color-coded based on the country's classification:

Developed countries: Displayed in red.

Developing countries: Displayed in blue.

From the plot, we observe that developed countries tend to have both higher GDP and higher Happiness Scores. The green dashed trend line suggests a positive correlation between GDP and Happiness Score, indicating that an increase in GDP may be associated with higher happiness levels.

Method Explanation

We use **quantile()** to calculate the 25th percentile (Q1) and 75th percentile (Q3) of GDP and determine the **interquartile range** (IQR).
Using these values, we define lower and upper bounds as Q1 - 1.5 × IQR and Q3 + 1.5 × IQR, filtering out GDP outliers to ensure extreme values do not distort the analysis.
Countries are then classified into developed or developing using the **apply** function.
Finally, we use **plt.scatter** to plot the GDP vs. Happiness Score for both groups. A first-degree polynomial regression (**np.polyfit**) is applied to fit a trend line, which is displayed as a green dashed line to highlight the general trend.

3. Libraries Used

The following Python libraries are utilized in this project:

geopandas
A library for handling geospatial data. Although not directly used in the provided code, it is often required when working with geographic datasets and country-based visualizations.

plotly
Used for interactive visualizations.
express: Used to create a choropleth map displaying the World Happiness Index, allowing for dynamic interactions such as tooltips and color scaling.
graph_objects: Used for detailed customizations of the map, including coastline, land, and ocean colors.
plotly.io: Used for managing figure rendering and display settings.

IPython.display
Used to render and display interactive Plotly figures within Jupyter Notebook environments.

pandas
Used for data manipulation and analysis. It was employed to read the CSV dataset, rename columns, filter and transform data, and perform grouping and filtering operations.

numpy

Used for numerical computations and mathematical operations. It was applied to calculate quartiles, perform linear fitting, handle arrays, and conduct other mathematical calculations.

matplotlib
Primarily used for data visualization, including:
pyplot: To create various types of plots, set figure sizes, add titles, axis labels, and annotations.
cm and **colors**: Utilized for color mapping and normalization to enhance the visual appeal of the plots.

seaborn
A statistical data visualization library based on matplotlib, used for creating aesthetically pleasing and informative plots such as heatmaps, pair plots, and box plots. It simplifies plot customization and improves readability.

scikit-learn
Used for machine learning modeling and data preprocessing. The project employed **RandomForestRegressor** to build a predictive model and **StandardScaler** to standardize features.

shap
A library for model interpretability and explainability. It was used to compute SHAP values, visualize feature importance, and generate SHAP summary plots to better understand the impact of different features on predictions.

4. Result Description and Findings

4.1 Overview

First, we use the World Happiness Index map to understand the global distribution of happiness index from a macro perspective, and then use the histogram of the average happiness index distribution to grasp the overall shape of the data. Next, display the top ten and bottom ten countries in terms of happiness index, highlighting the two extreme situations. Afterwards, through correlation heatmaps, feature importance maps, and SHAP value summary maps, we will conduct in-depth analysis of the factors that affect the happiness index, their importance, and ways of influence. Further explore the relationships and data distribution characteristics between variables using a multi factor scatter plot matrix and a box plot of happiness related indicators. Finally, use a multi indicator line chart ranked by country to observe the trend of key indicators changing with the ranking, and comprehensively analyze the data.

4.2 Results

1. **The distribution of happiness index is uneven globally**. Countries in Western Europe and Oceania have generally higher happiness index due to their developed economy and good social welfare, while some countries in Africa and South Asia have lower happiness index. The distribution of the average happiness index follows a mode of around 6, resembling a unimodal bell shaped curve, but it is asymmetric. The distribution of countries in low happiness index regions is scattered, while the number of countries in high happiness index regions is small, reflecting significant differences in happiness index among countries.

2. **Explore the influencing factors of happiness index**. The correlation heatmap shows a strong positive correlation between happiness index and social support, GDP and healthy life expectancy, indicating that these factors have a significant impact on happiness index. The feature importance bar chart of the random forest model clarifies the importance ranking of factors that affect the happiness index, with **social support, GDP, and other factors decreasing in order**. The SHAP summary once again confirms that. The multivariate scatter plot matrix reveals the happiness index GDP The pairwise relationship between social support and healthy life expectancy is most concentrated in the moderate happiness category among these four indicators.

3. **The relationship between various factors and happiness index is complex**. The box plot of happiness related indicators shows that the overall scores of GDP and social support are high and scattered, while the scores of perceived corruption and generosity are low and concentrated, and all indicators have relatively symmetrical distributions. The multi indicator trend chart shows that the happiness index curve is steep at high and low rankings, flat in the middle, while GDP, social support, and healthy life expectancy have no significant correlation with rankings and fluctuate greatly in low ranking countries. Developed countries often have higher GDP and happiness index, and GDP is positively correlated with happiness index, that is, GDP growth may bring higher levels of happiness.

4.3 Insights from the Chosen Dataset

1. **The key role of social support**: Social support is **the primary factor** affecting happiness in 2024, highlighting the central position of social environment and interpersonal relationships in people's happiness. If a country or region can provide a strong social support system for its people, such as comprehensive social security and a good community atmosphere, it will greatly enhance their sense of happiness.

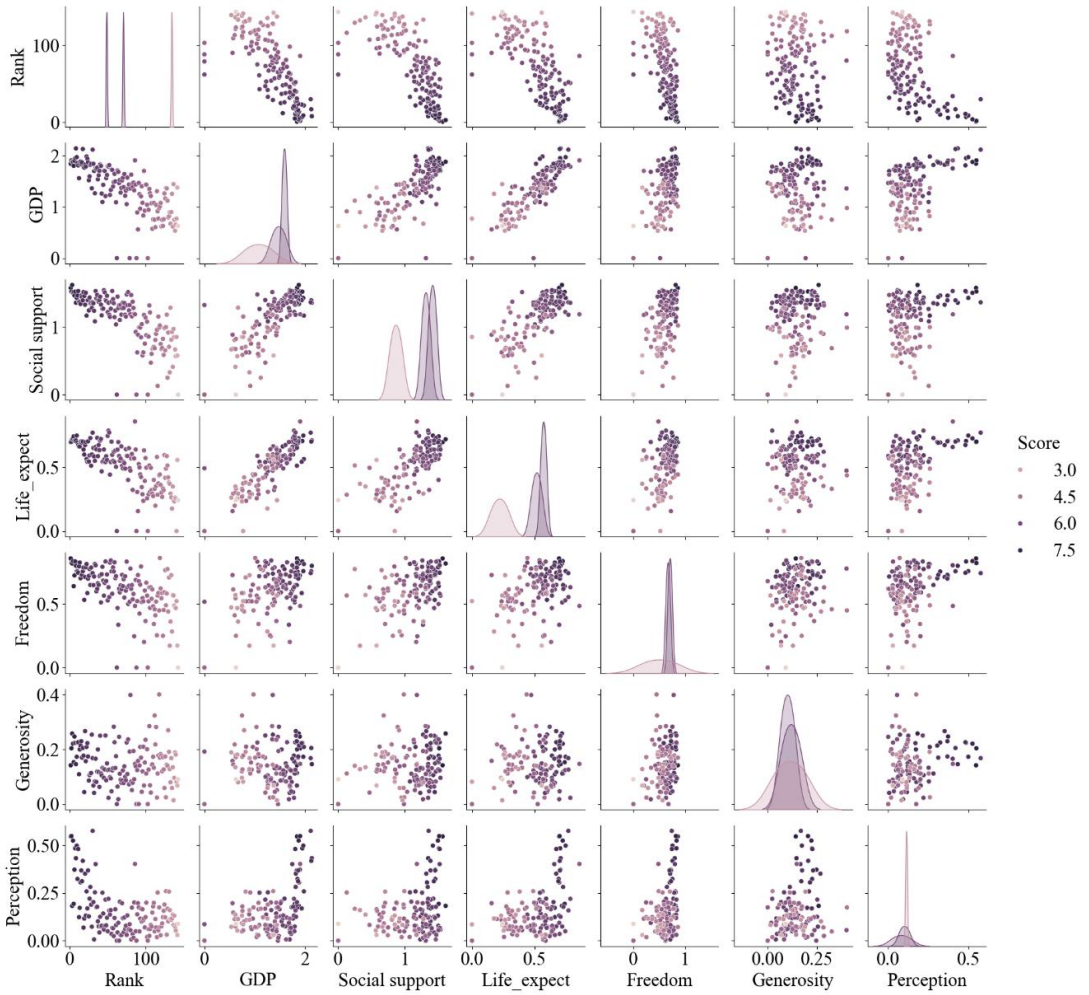
2. **The importance of economic factors**: GDP and other economic factors are second only to social support in affecting happiness, as they can provide more resources for society to improve education, healthcare, infrastructure, and indirectly enhance

people's quality of life and happiness.

3. **The interdependence of various factors**: These factors do not exist in isolation, but are **interrelated and influence each other**. For example, economic development may promote the improvement of the social support system, and good social support may further promote stable economic growth. Understanding the relationship between these factors can help formulate more comprehensive and effective policies to comprehensively enhance people's sense of happiness.

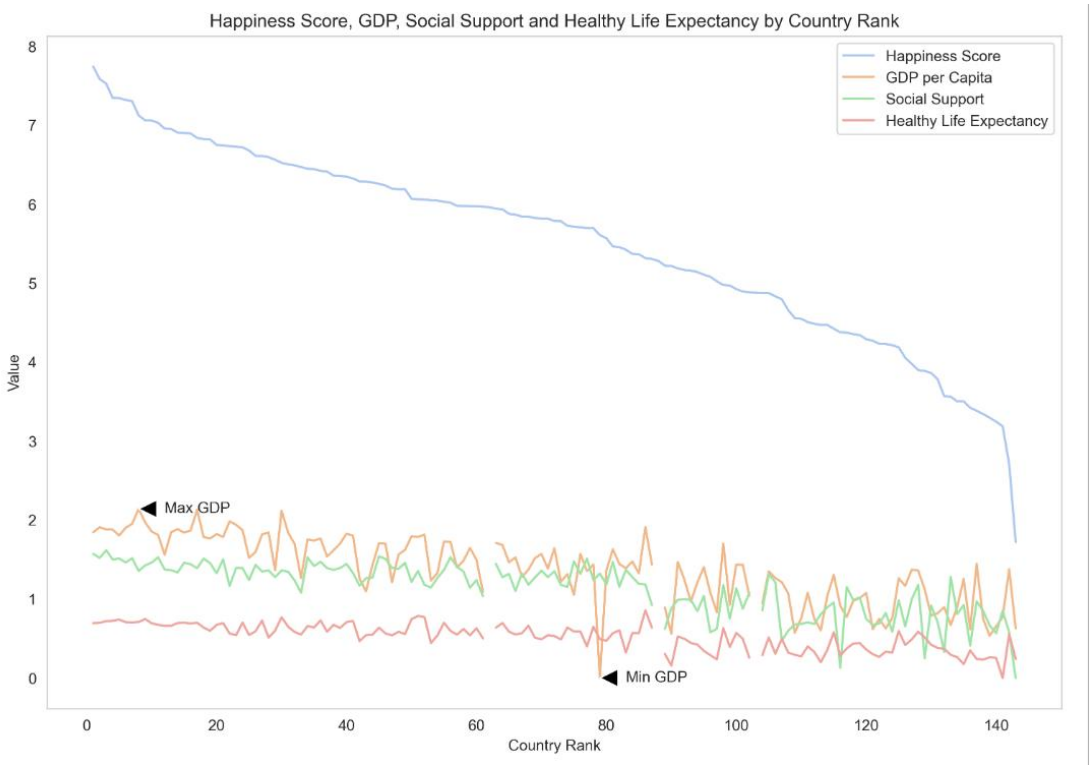
According to the comprehensive analysis results, various factors such as social support, economic development, and healthy life expectancy jointly affect people's sense of happiness. The effective coordination and promotion of these factors require good governance. So the government and society should attach importance to enhancing governance capabilities to meet the people's pursuit of a happy life.

Appendix



This chart can assist in demonstrating the complex relationship between influencing factors and happiness scores. Social support and economic conditions may not necessarily have a simple linear relationship with happiness scores.

```
sns.pairplot(data=df, hue='Score')
```



This is a line chart made without data preprocessing, which can illustrate the special situations of a very small number of countries.