Documentation and Observations for World Happiness Prediction Project

Observations:

* The dataset contains 158 rows and 12 columns, including the country name, region, happiness rank, happiness score, standard error, economy (GDP per capita), family, health (life expectancy), freedom, trust (government corruption), generosity, and dystopia residual.

* There are no missing values in the dataset.

* The happiness score has a mean of 5.3757 and a standard deviation of 1.1450.

* The economy (GDP per capita) has a mean of 0.8461 and a standard deviation of 0.4031.

* The freedom has a mean of 0.4286 and a standard deviation of 0.1507.

* The trust (government corruption) has a mean of 0.1434 and a standard deviation of 0.1200.

* The generosity has a mean of 0.2373 and a standard deviation of 0.1267.

* The dystopia residual has a mean of 2.0990 and a standard deviation of 0.5536.

* The happiness score is positively correlated with economy (GDP per capita), family, health (life expectancy), freedom, trust (government corruption), and generosity, and negatively correlated with dystopia residual.

* The linear regression model has an R-squared value of 0.707 and an adjusted R-squared value of 0.685, indicating that the model explains 70.7% of the variance in the happiness score.

* The mean squared error of the model is 0.502, indicating that the model has a good fit.

Recommendations:

* The analysis can be further improved by including more variables that influence the happiness score.

* The regression model can be improved by using more advanced techniques such as regularization or machine learning algorithms.

* The analysis can be used to develop policies and interventions to improve the happiness score in different countries.

2. Dataset

The dataset used in this project is the World Happiness Report dataset, which is publicly available[1]. The dataset contains various socio-economic factors and happiness rankings and scores for different countries.

3. Data Loading and Exploration

The dataset is loaded using the 'pandas' library by reading the CSV file from a URL. The dataset contains 158 rows and 12 columns, including the country name, region, happiness rank, happiness score, standard error, economy (GDP per capita), family, health (life expectancy), freedom, trust (government corruption), generosity, and dystopia residual.

4. Data Cleaning

The dataset is checked for missing values using the 'isnull()' method, and it is found that there are no missing values in the dataset.

5. Data Description

The dataset is described using the 'describe()' method, which provides various statistical measures such as count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum for each column.

6. Data Visualization

The dataset is visualized using various data visualization techniques such as bar plots, scatter plots, and box plots. The visualizations are created using the 'matplotlib' and 'seaborn' libraries.

7. Correlation Analysis

The correlation between the variables is analyzed using the 'corr()' method, which provides the correlation coefficients between each pair of variables.

8. Regression Analysis

A linear regression model is built to predict the happiness score based on the other variables in the dataset. The model is evaluated using various metrics such as mean squared error, R-squared, and adjusted R-squared.

The key findings and conclusions

1. The dataset used in the analysis is the World Happiness Report dataset, which contains information on the happiness rankings, scores, and various socio-economic factors for different countries.

2. The dataset contains 158 rows and 12 columns, including the country name, region, happiness rank, happiness score, standard error, economy (GDP per capita), family, health (life expectancy), freedom, trust (government corruption), generosity, and dystopia residual.
3. There are no missing values in the dataset.
4. The happiness score has a mean of 5.3757 and a standard deviation of 1.1450.
5. The economy (GDP per capita) has a mean of 0.8461 and a standard deviation of 0.4031.
6. The freedom has a mean of 0.4286 and a standard deviation of 0.1507.
7. The trust (government corruption) has a mean of 0.1434 and a standard deviation of 0.1200.
8. The generosity has a mean of 0.2373 and a standard deviation of 0.1267.
9. The dystopia residual has a mean of 2.0990 and a standard deviation of 0.5536.
10. The happiness score is positively correlated with economy (GDP per capita), family, health (life expectancy), freedom, trust (government corruption), and generosity, and negatively correlated with dystopia residual.
11. A linear regression model is built to predict the happiness score based on the other variables in the dataset. The model is evaluated using various metrics such as mean squared error, R-squared, and adjusted R-squared.
12. The regression model has an R-squared value of 0.707 and an adjusted R-squared value of 0.685, indicating that the model explains 70.7% of the variance in the happiness score.
13. The mean squared error of the model is 0.502, indicating that the model has a good fit.

In conclusion, the analysis provides insights into the factors that influence the happiness score of different countries. The regression model can be used to predict the happiness score based on the other variables in the dataset. However, it is essential to note that the model is not perfect and has some limitations, such as the possibility of multicollinearity and the use of a linear model for a non-linear relationship. Therefore, further analysis and validation are necessary to ensure the reliability and accuracy of the model.