**ESIEE Paris**



**E4\_AIC\_4301B Data Science Project**

**World Happiness Report**

**Mehmet Hasat SERINKAN**

**Efe ALTIN**

**Course Instructor: Gaelle CANDEL**

Link for the dataset: https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2023

**Context**

**1. Introduction**

**1.1 Why We Choose This Dataset**

**1.2 Describing the Data**

**1.3 Why it is Complex**

**1.4 Describing the Task**

**2. Analysis**

**2.1 Data Preparation**

**2.2 Graphs**

**2.3 Applying ML Algorithm**

**3. Reference**

**1) Introduction**

**1.1) Why We Choose This Dataset**

1. Rich Data Source: The World Happiness Report 2023 dataset contains a vast amount of data related to various aspects of happiness, including Logged GDP per Capita, Social Support, Healthy Life Expectancy, Genorosity and Perceptions of Corruption, which can be analyzed and explored through various data science techniques. The dataset also includes data on different countries which providing a broad range of perspectives to explore
2. Relevance: Happiness is a core human goal Therefore, analyzing data from the World Happiness Report can provide valuable insights into the factors that contribute to happiness and inform policies aimed at improving people's quality of life.
3. Practical Applications: this report can also be useful in practice. For example, social welfare organizations can use the findings to design interventions that support vulnerable populations.

**1.2) Describing the Data**

There are 13 columns in this dataset and these are:

1. Country Name
2. Ladder Score
3. Standart Error of Ladder Score
4. Upperwhisker
5. Lowerwhisker
6. Logged GDP per Capita
7. Social Support
8. Healthy Life Expectancy
9. Freedom to Make Life Choices
10. Generosity
11. Perceptions of Corruption
12. Ladder Score in Dystopia
13. Dystopia + Residual

**1.3) Why it is Complex**

The dataset used for the World Happiness Report includes a wide range of variables that are collected from different sources, such as international surveys and official statistics from governments and international organizations. The variables cover various aspects of life, such as social support, freedom to make life choices, and perceptions of corruption. One of the main challenges of this dataset is that it contains both objective and subjective measures of happiness and well-being. For example, income and life expectancy are objective measures, while self-reported measures of happiness and life satisfaction are subjective. Combining these different types of measures and ensuring their validity and reliability can be a complex task. Moreover, the dataset covers a large number of countries. the data may not be consistent or comparable across different countries. Therefore, preprocessing the data to ensure its quality and comparability is an essential step in analyzing this dataset.

**1.4) Describing the Task**

The task of this project is to perform data analysis on the World Happiness Report 2023 dataset to gain insights and understanding of the factors that contribute to the happiness of different countries.

The project aims to explore the dataset using various data analysis techniques such as data cleaning, data visualization, statistical analysis. The analysis is performed using Python programming language and various libraries such as pandas, seaborn, and matplotlib.

The project also aims to develop a predictive model to predict the happiness score of different countries based on the available features. This is done using a linear regression model and the model performance is evaluated using various metrics such as R-squared and mean absolute error.

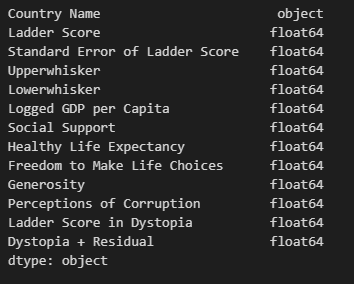
**2) Analysis**

**2.1) Data Preparation**

The World Happiness Report 2023 dataset contains 137 rows and 19 columns. However, we dropped some unnecessary columns and renamed the columns.

Table 1: Description of the data



 Table 2: Count Table 3: Data dtypes

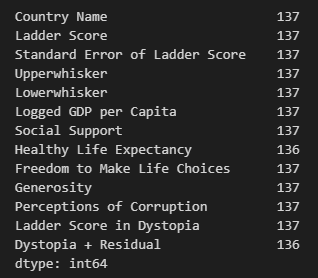


Table 4: Info

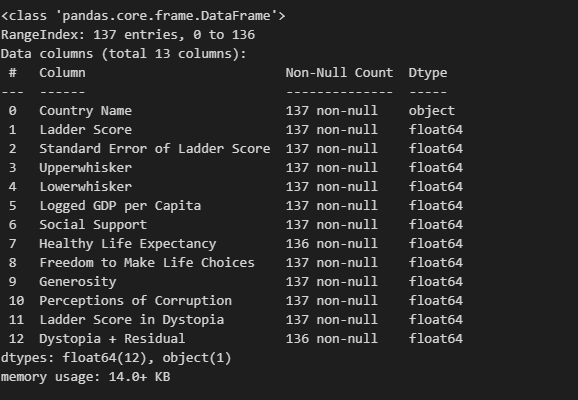
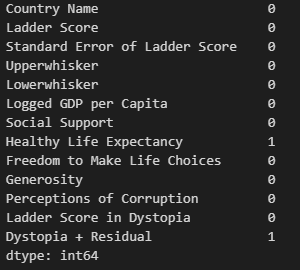


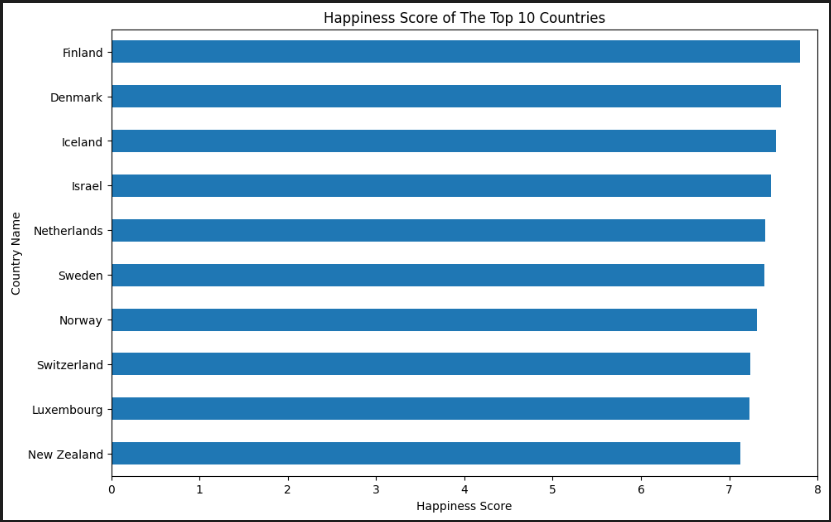
Table 5: Total Number of Nulls



**2.2) Graphs**

We created some graphs to understand the data. These are some examples.

Table 6: Happiness Score of the top 10 countries

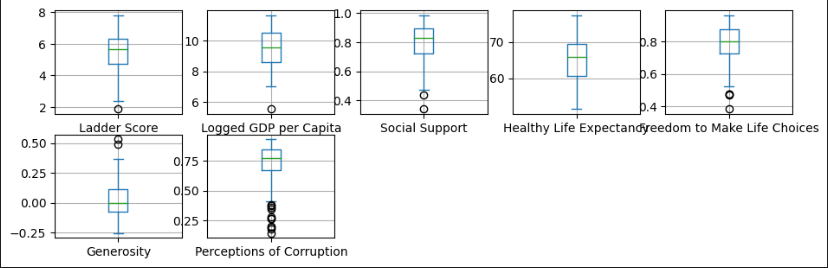


First, the data DataFrame is indexed by the Country Name column using the set\_index() method, and then the first 10 rows of the Ladder Score column are selected using data\_plot[0:10]['Ladder Score']. This creates a new DataFrame that only includes the Ladder Score values for the top 10 countries.

Next, a figure with a size of 11x7 inches is created using the subplots() function from the matplotlib library, and the resulting figure and axis objects are stored in the fig and ax variables, respectively.

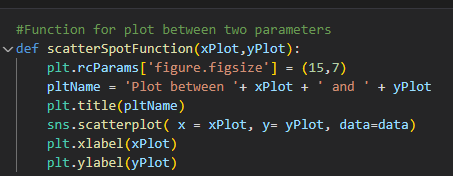
Then, a horizontal bar plot is created using the plot.barh() method on the data\_plot DataFrame, which creates horizontal bars for each country's Ladder Score.

Table 7: Distribution and Spread of each variable



Creating a grid of box plots to visualize the distribution of the seven variables in the dataset: Ladder Score, Logged GDP per Capita, Social Support, Healthy Life Expectancy, Freedom to Make Life Choices, Generosity, and Perceptions of Corruption. The resulting plot shows a grid of box plots, with one box plot for each variable. The box plots display the distribution of each variable.

Table 8: scatterSpotFunction



This code defines a function called scatterSpotFunction() that creates a scatter plot between two parameters, xPlot and yPlot, using the Seaborn library.

Table 9: Ladder Score and Logged GDP per Capita

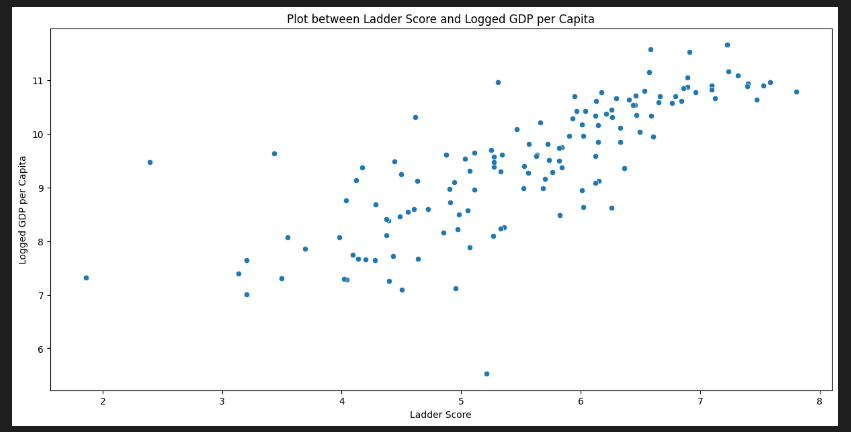


Table 10: Ladder Score and Social Support

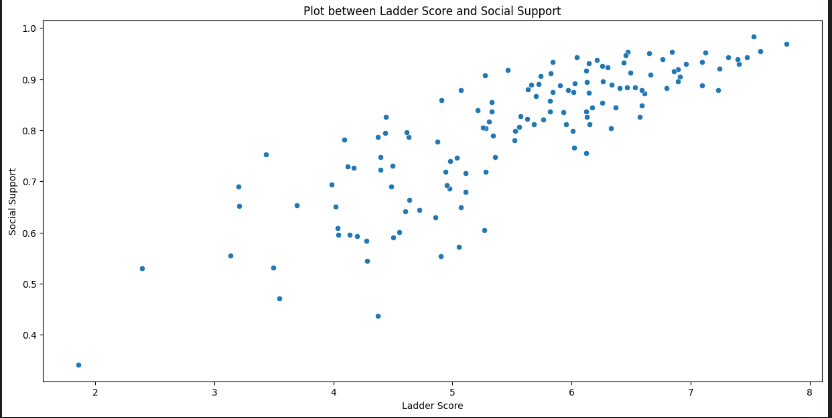


Table 11: Ladder Score and Healthy Life Expectancy



Table 12: Ladder Score and Freedom to Make Life Choices

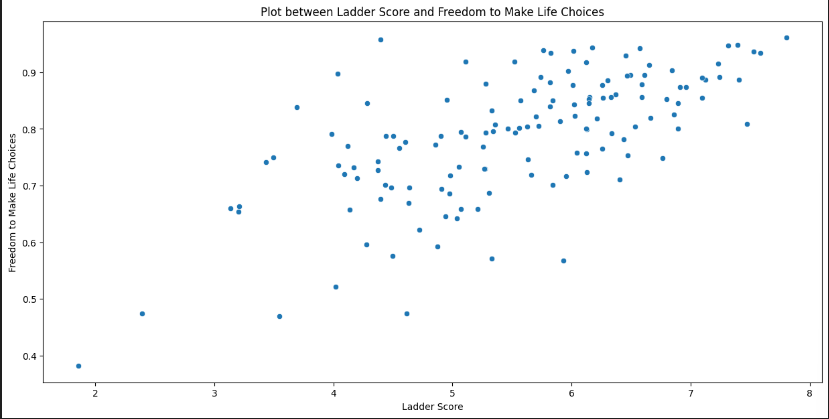


Table 13: Ladder Score and Generosity

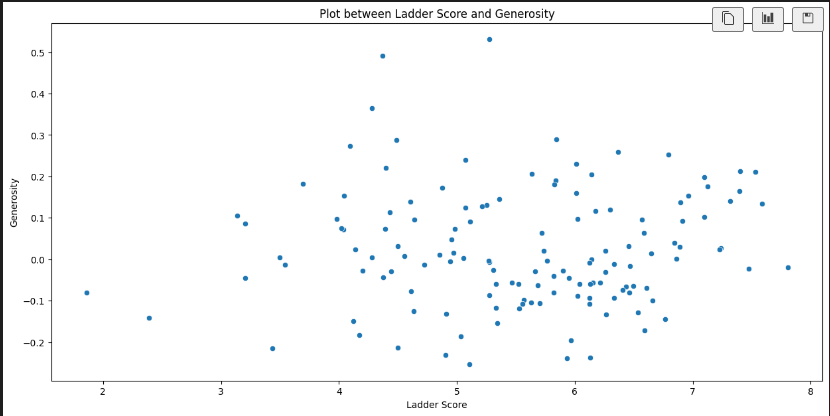


Table 14: Ladder Score and Perceptions of Corruption

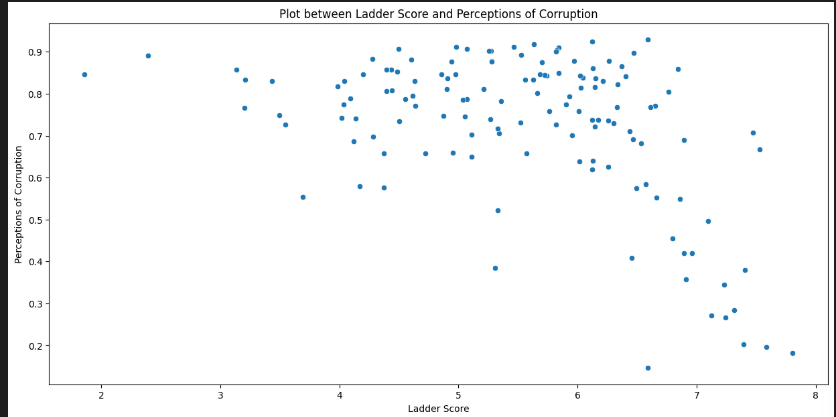


Table 15: Correlation between the Happiness Score and each of the other variables

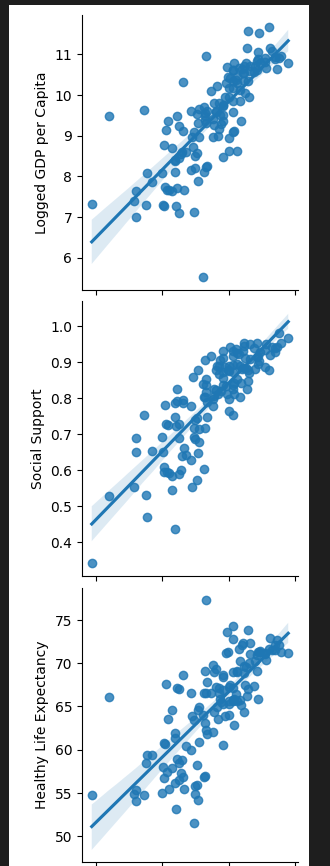
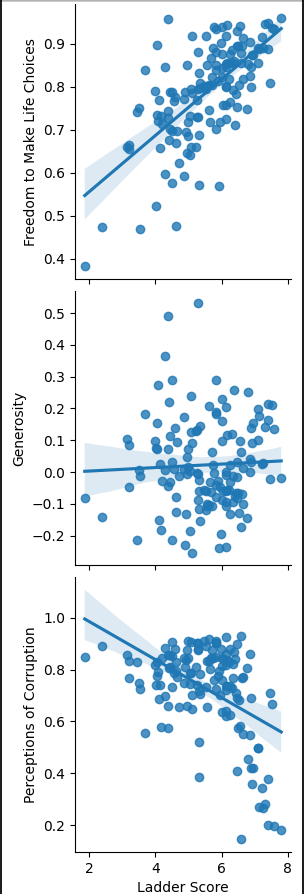


Table 16: Part 2

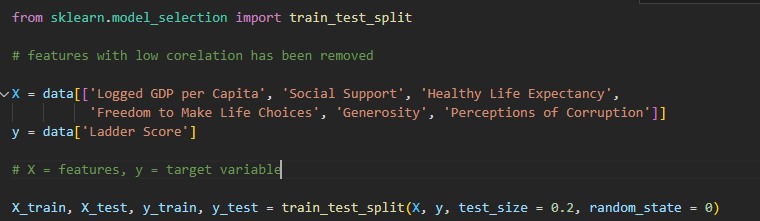


In each scatter plot, the Ladder Score variable is plotted on the x-axis, and the variable specified in y\_vars is plotted on the y-axis. The kind parameter is set to 'reg', which adds a regression line to each scatter plot, showing the trend in the data.

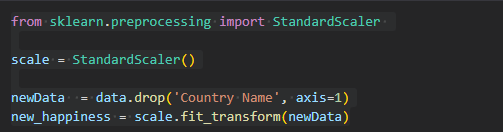
The resulting grid of scatter plots allows us to visualize the relationships between the Ladder Score variable and each of the other variables, and to assess the strength and direction of these relationships. The regression lines give us a sense of the overall trend in the data, while the scatter plots allow us to see how much variation there is in the data and whether there are any outliers or unusual data points.

Overall, this type of plot is useful for exploring the relationships between multiple variables and identifying any potential patterns or trends in the data.

**2.3) Applying ML Algorithm**

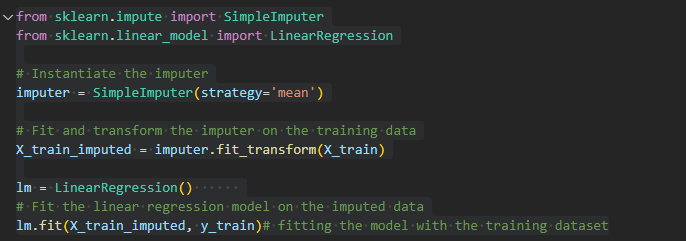


This code is performing the data preprocessing step for building a predictive model. It is splitting the data into training and testing sets using the train\_test\_split() function from the sklearn.model\_selection module. The test\_size parameter is set to 0.2, which means that 20% of the data will be used for testing and the remaining 80% will be used for training. The random\_state parameter is set to 0 to ensure that the data is split in the same way every time the code is executed.



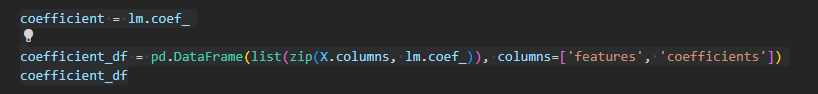
This code performs feature scaling on the dataset using the StandardScaler method from the sklearn.preprocessing library. In this code, the StandardScaler object is instantiated and stored in the variable scale. Then, the Country Name column is dropped from the original data dataframe, and the resulting dataframe is stored in the variable newData.

The fit\_transform method of the StandardScaler object is then used to scale the features in newData, which is stored in the variable new\_happiness. The resulting new\_happiness array contains the scaled values of all the features in the dataset except for Country Name.



Imputing missing values: First, it uses the SimpleImputer from sklearn.impute module to replace missing values with the mean value of the respective feature. This is done to make sure that the model is not biased towards the missing values and can still learn from the other non-missing values.

Fitting the Linear Regression model: After imputing the missing values, it then fits a Linear Regression model on the training dataset using the fit() method from the LinearRegression module. This creates a model that can predict the target variable (in this case, the Ladder Score) based on the input features.



This code computes the coefficients of the linear regression model and creates a dataframe to store them with the corresponding features.

Result: The resulting dataframe coefficient\_df displays the coefficients of the linear regression model for each feature.



The result shows the coefficients of the linear regression model, which indicates how much the Ladder Score changes when a feature is increased by one unit, while holding all other features constant.

Logged GDP per Capita has a coefficient of 0.150286, which means that a one-unit increase in the Logged GDP per Capita feature is associated with an increase of 0.150286 units in the Ladder Score, holding all other features constant.

Social Support has the highest coefficient of 4.157277, which means that a one-unit increase in the Social Support feature is associated with an increase of 4.157277 units in the Ladder Score, holding all other features constant.

Healthy Life Expectancy has a coefficient of 0.030778, which means that a one-unit increase in the Healthy Life Expectancy feature is associated with an increase of 0.030778 units in the Ladder Score, holding all other features constant.

Freedom to Make Life Choices has a coefficient of 2.437461, which means that a one-unit increase in the Freedom to Make Life Choices feature is associated with an increase of 2.437461 units in the Ladder Score, holding all other features constant.

Generosity has a coefficient of 0.040156, which means that a one-unit increase in the Generosity feature is associated with an increase of 0.040156 units in the Ladder Score, holding all other features constant.

Perceptions of Corruption has a coefficient of -0.781299, which means that a one-unit increase in the Perceptions of Corruption feature is associated with a decrease of 0.781299 units in the Ladder Score, holding all other features constant.

Overall, the features of Social Support, Freedom to Make Life Choices, and Logged GDP per Capita have the highest positive impact on Ladder Score, while Perceptions of Corruption has the highest negative impact on Ladder Score.

**3. Reference**

[**https://www.youtube.com/watch?v=kLDTbavcmd0&list=LL&index=4&t=2499s**](https://www.youtube.com/watch?v=kLDTbavcmd0&list=LL&index=4&t=2499s)

[**https://www.youtube.com/watch?v=xi0vhXFPegw&list=LL&index=6&t=46s**](https://www.youtube.com/watch?v=xi0vhXFPegw&list=LL&index=6&t=46s)

**https://www.youtube.com/watch?v=HNtEq-dK3C4&list=LL&index=3**