Name: أحمد مجاهد محمود حسن الجمال

ID: 2051122212

Question 2:

Summary of the World Happiness Report 2015 Dataset

The dataset consists of a single CSV file with 12 columns:

1. Country: the name of the country or region being evaluated.
2. Region: the region in which the country or region is located.
3. Happiness Rank: the ranking of the country or region based on the Happiness Score.
4. Happiness Score: a metric calculated based on the answers to the Gallup World Poll, to evaluate their overall life satisfaction on a scale of 0 to 10.
5. Standard Error: the standard error of the Happiness Score.
6. GDP per capita: the country's economic productivity based on its Gross Domestic Product (GDP) per capita.
7. Family: the social support score for the country or region.
8. Life Expectancy: the healthy life expectancy score for the country or region.
9. Freedom: the score measuring freedom to make life choices.
10. Trust: the perception of corruption score for the country or region.
11. Generosity: the generosity score for the country or region.
12. Dystopia Residual: a calculated score based on the values of the other columns in the dataset.

The following observations regarding its quality:

1. **Validity**: The data seems to be valid as it is sourced from reputable organizations and has been used in various studies and reports.
2. **Accuracy**: The accuracy of the data cannot be determined without further information on the methodology used to collect the data.
3. **Completeness**: The data is complete; we cannot assume that there are no missing values in other columns without further investigation.
4. **Consistency**: The data seems to be consistent as there are no major inconsistencies or contradictions between the columns.
5. **Uniformity**: The data is uniform for all columns with the right format.

Based on the data exploration and analysis conducted, the following types of data dirtiness were identified:

1. **Missing Data**: There were no missing values in the dataset.
2. **Outliers**: Outliers were detected in some variables such as Standard Error, Family, Trust (Government Corruption), Generosity, and Dystopia Residual. Outliers can affect the mean and standard deviation of a variable, which can, in turn, affect the accuracy of statistical models.
3. **Inconsistencies**: No major inconsistencies were found in the data.
4. **Inaccurate Data**: There were no major inaccuracies in the data.
5. **Incomplete Data**: There was no major incomplete data in the dataset.

Question 3:

Exploring the World Happiness Report 2015: Key Findings

Bird View on the data

The data contains information on 158 countries across 12 columns.

Our findings show that there were no missing values or duplicates in the dataset.

The countries were ranked based on their happiness score, which was calculated using various factors such as GDP per capita, life expectancy, and government corruption.

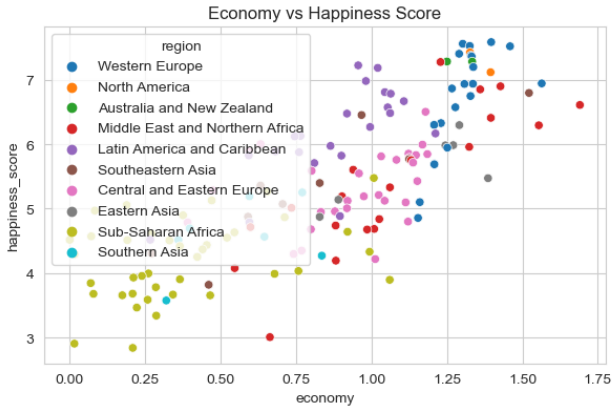
Analyzing the distributions of all variables

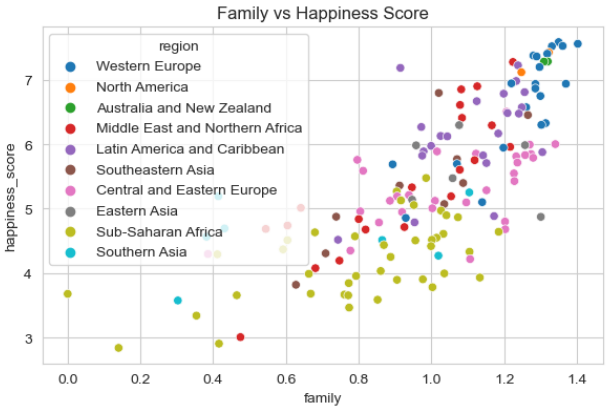
- **Happiness score**: The mean is greater than the median, which indicates that the distribution of happiness scores is skewed to the right. This means that there are more countries with higher happiness scores than lower scores.
- **Economy, Family, Health, Generosity, and Dystopia**: The median is greater than the mean for most of these variables, which suggests that the distributions are skewed to the left. This means that there are more countries with lower values of these variables than higher values.
- **Correlation between happiness score and other variables**: The highest correlation is between happiness score and economy, which suggests that a strong economy is strongly associated with high happiness scores. There is also a strong positive correlation between happiness score and family, health, and dystopia. However, the correlation between happiness score and freedom is weaker.

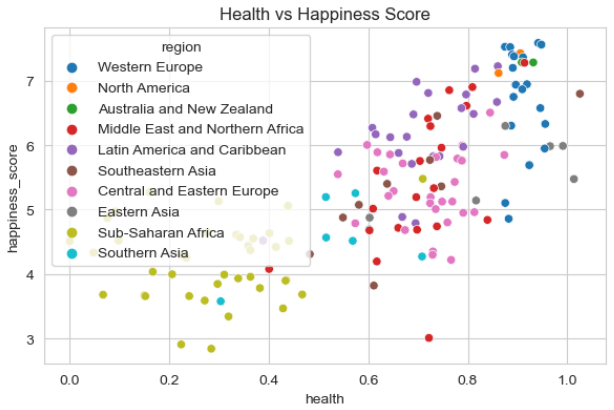
The benefits of this:

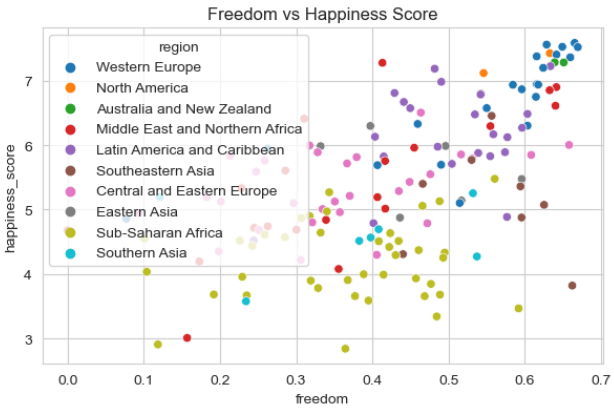
- Countries with strong economies tend to have higher happiness scores.
- Countries with strong family ties tend to have higher happiness scores.
- Countries with good health care tend to have higher happiness scores.
- Countries with low levels of dystopia tend to have higher happiness scores.
- Freedom is not as important for happiness as other factors, such as economy, family, and health.

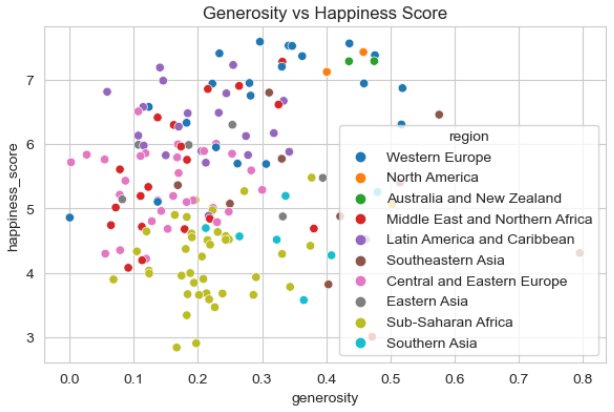
The Effect of region on happiness score

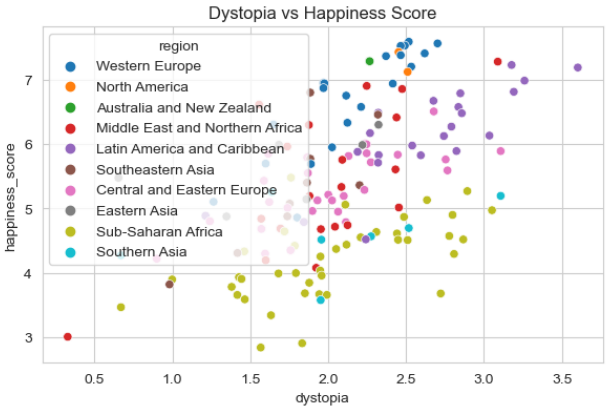


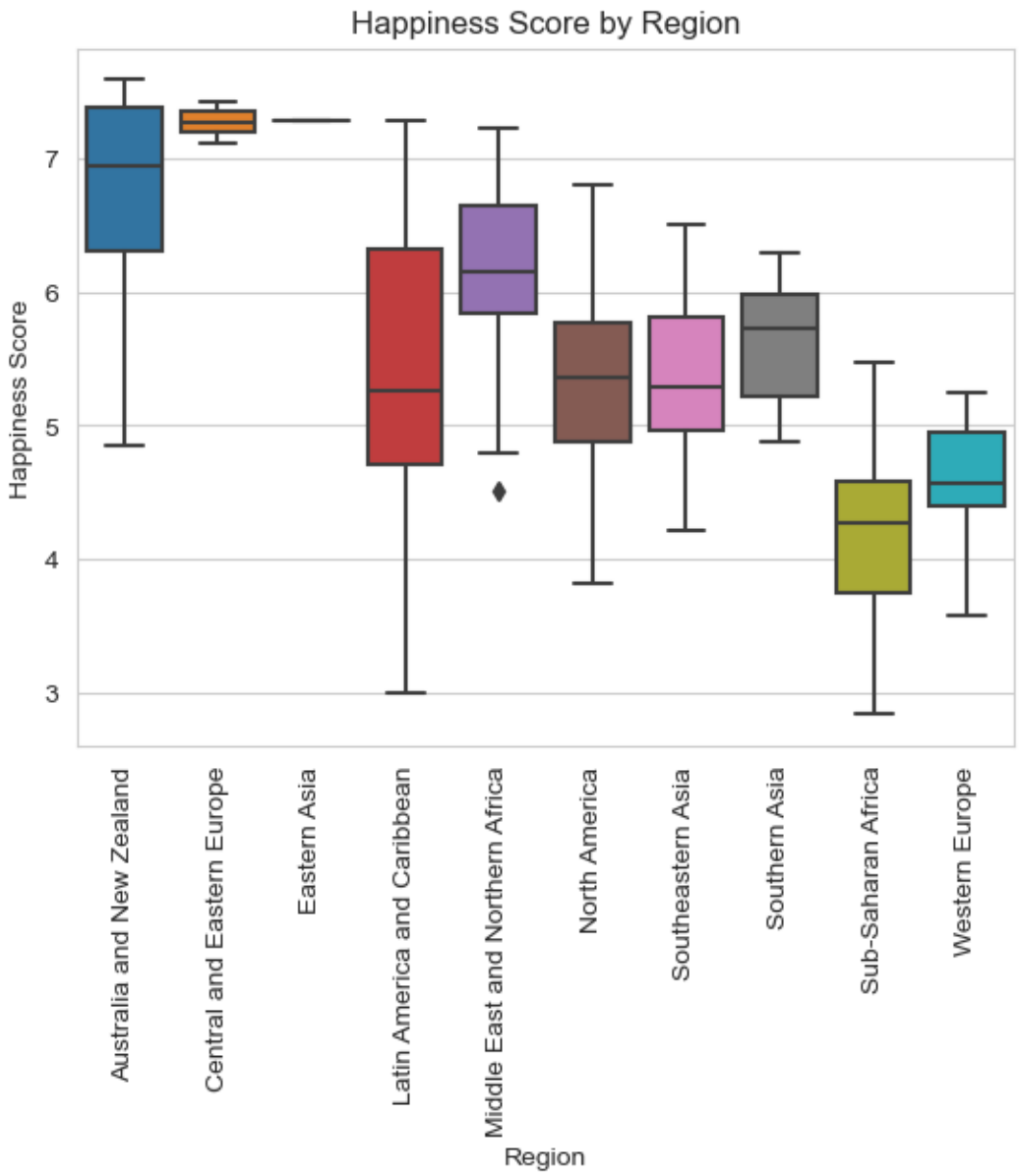












- The data shows that there are no outliers in most regions, except for Latin America and the Caribbean.

- The boxplots also show that the distribution of happiness scores is more normal in North America and Central and Eastern Europe.

- The scatterplots show that there is a strong correlation between happiness score and economy in all regions. This suggests that a strong economy is strongly associated with high happiness scores in all regions.

- There is also a strong positive correlation between happiness score and family, health, and dystopia in all regions.

- The correlation between happiness score and freedom is weaker in some regions, such as Sub-Saharan Africa and Southern Asia. This suggests that freedom is not as important for happiness in these regions as other factors, such as economy, family, and health.

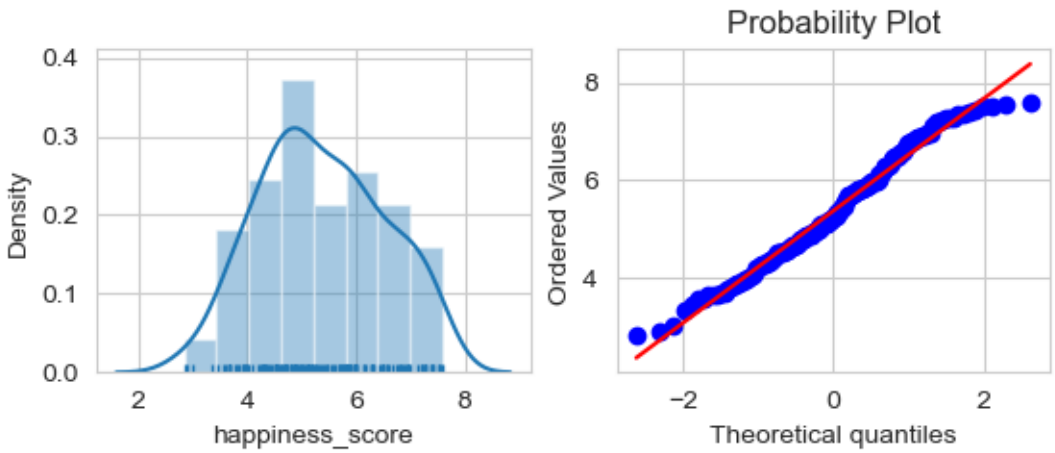
The suggests that came out of this:

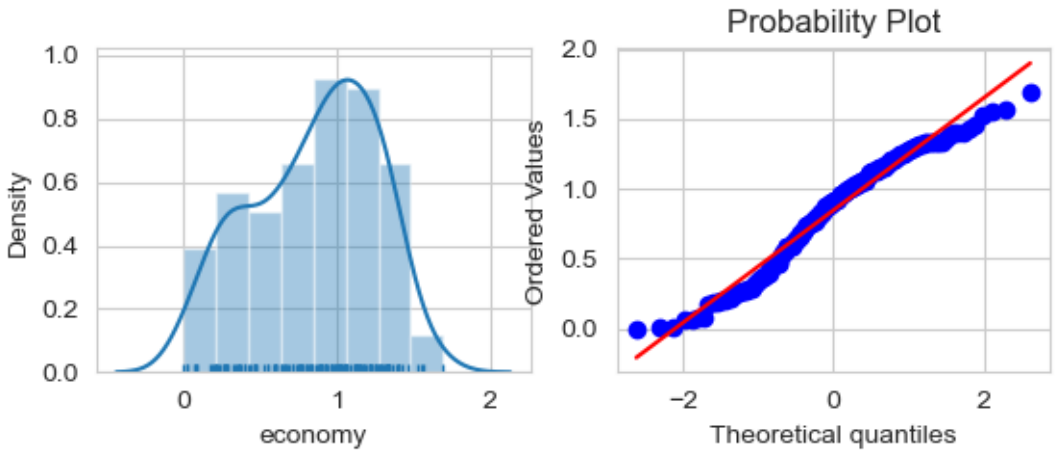
- This suggests that the distribution of happiness scores is more evenly distributed in most regions.
- This suggests that most people in these regions have similar happiness scores, while there is more variation in happiness scores in other regions.
- This suggests that a strong economy is strongly associated with high happiness scores in all regions.
- This suggests that strong family ties, good health, and a low level of dystopia are all associated with high happiness scores in all regions.

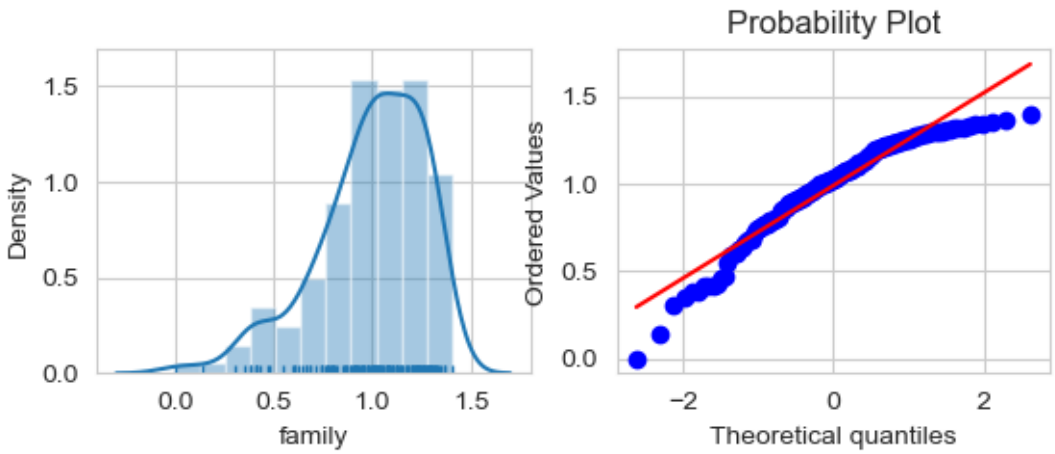
- This suggests that freedom is not as important for happiness in these regions as other factors, such as economy, family, and health.

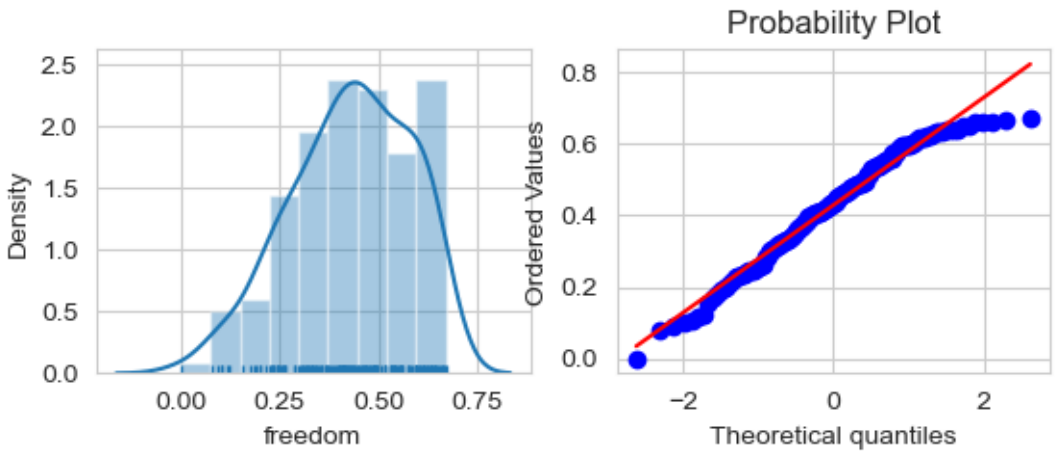
Analyzing for Normalization test

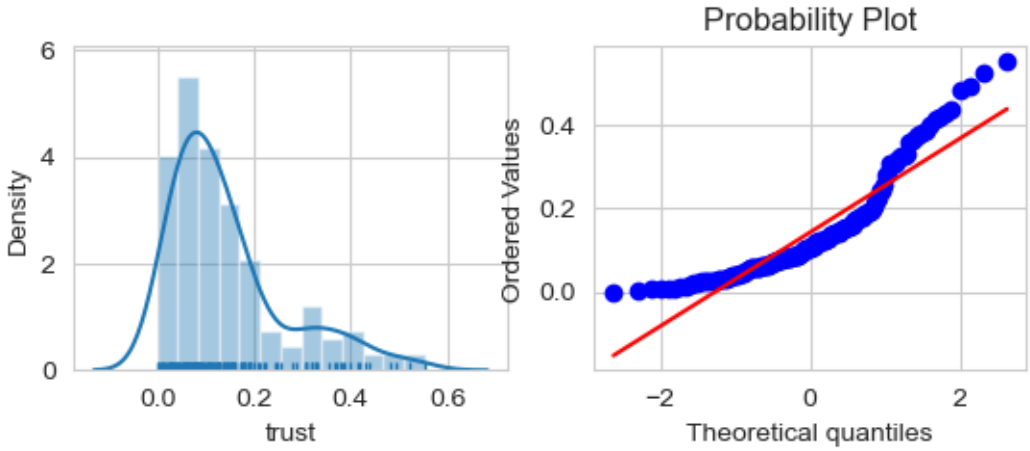
The distribution for each variable

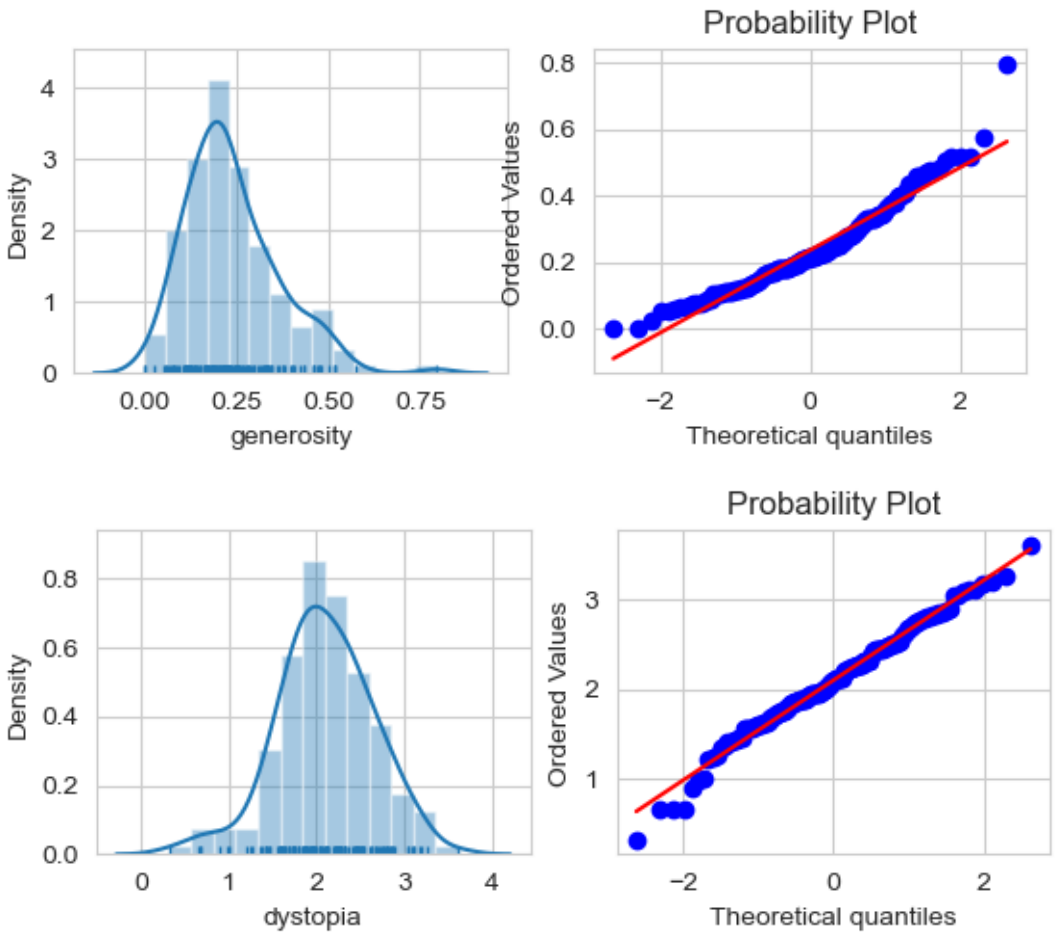







Density
freedom
Probability Plot
Ordered Values
Theoretical quantiles


Density
trust
Probability Plot
Ordered Values
Theoretical quantiles



Note: The Shapiro–Wilk test is a normality test that is used to determine if a data set is normally distributed. The test statistic is a measure of how well the data fits a normal distribution.

By applying Shapiro-Wilk test:

Based on the output, we can see that all variables except for dystopia have p-values less than 0.05, indicating that we reject the null hypothesis and conclude that the distribution of each variable is not normal. This means that these variables do not follow a normal distribution, and any analysis or modeling assumptions that rely on normality may not be valid.

Based on the observations from the plots:

- Happiness score has a positive skewness as its mean is greater than the median, indicating that there are some countries with extremely high happiness scores that are pulling the mean up.
- Economy has a negative skewness as its median is greater than the mean, indicating that there are some countries with very low economy scores that are pulling the median down.
- Family has a negative skewness as its median is greater than the mean, indicating that there are some countries with very low family scores that are pulling the median down.
- Health has a negative skewness as its median is greater than the mean, indicating that there are some countries with very low health scores that are pulling the median down.
- Freedom has a roughly symmetrical distribution as its mean is close to the median.
- Generosity has a positive skewness as its mean is greater than the median, indicating that there are some countries with extremely high generosity scores that are pulling the mean up.
- Dystopia has a roughly symmetrical distribution as its mean is close to the median.

Apply Linear Regression analysis.

The output of predicting:

R-squared: 0.9922
Adjusted R-squared: 0.9919
RMSE: 0.1144
MAE: 0.0801

Based on output of applying linear regression analysis the following observations can be made:

- The independent variables included in the model (economy, family, health, freedom, generosity, dystopia residual) are all factors that have been shown to contribute to happiness in previous research.
- The model has a high R-squared value of 0.9922, indicating that the independent variables in the model (economy, family, health, freedom, generosity, dystopia residual) explain a large proportion of the variation in the dependent variable (happiness score).
- The adjusted R-squared value of 0.9919 suggests that the independent variables in the model are relevant and contribute significantly to the model's predictive ability, even after accounting for the number of variables included in the model.
- The RMSE of 0.1144 indicates that the model's predictions have an average error of 0.1144 units. This means that, on average, the predicted happiness score differs from the actual happiness score by about 0.1144 units.
- The MAE of 0.0801 indicates that, on average, the model's predictions differ from the actual happiness score by 0.0801 units, which is a relatively small error.

Evaluating the Performance of Linear Regression Model on World Happiness Report 2016 Dataset

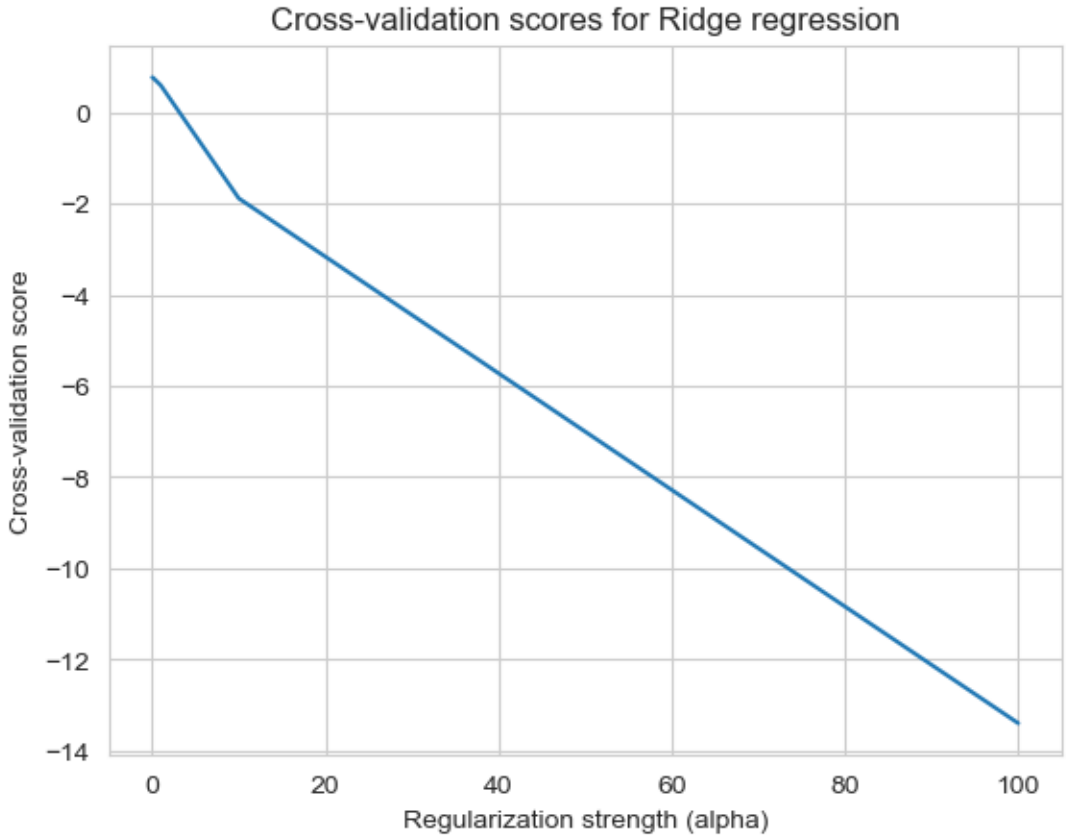
1) Test new data from 2016 version.

Test this random sample of the data

	Country	Happiness Score	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
21	Singapore	6.739	1.64555	0.86758	0.94719	0.48770	0.46987	0.32706	1.99375
70	Romania	5.528	1.16970	0.72803	0.67602	0.36712	0.00679	0.12889	2.45184
142	South Sudan	3.832	0.39394	0.18519	0.15781	0.19662	0.13015	0.25899	2.50929

Country	Real Value	Predicted values
Singapore	6.739	6.48006533
Romania	5.528	5.61467206
South Sudan	3.832	4.0762715

2) Evaluated the Linear Regression model using Regularization & Cross-validation



Summary

Based on the cross-validation scores, it appears that the model is performing reasonably well and is not overfitting. The cross-validation scores range from 0.60 to 0.91, with a mean score of 0.76. This indicates that the model is able to generalize well to new data.

Clustering the data using K-Means

The features are [economy, family, health, freedom, generosity, dystopia]

Apply Standard Scaler on subset of the data.

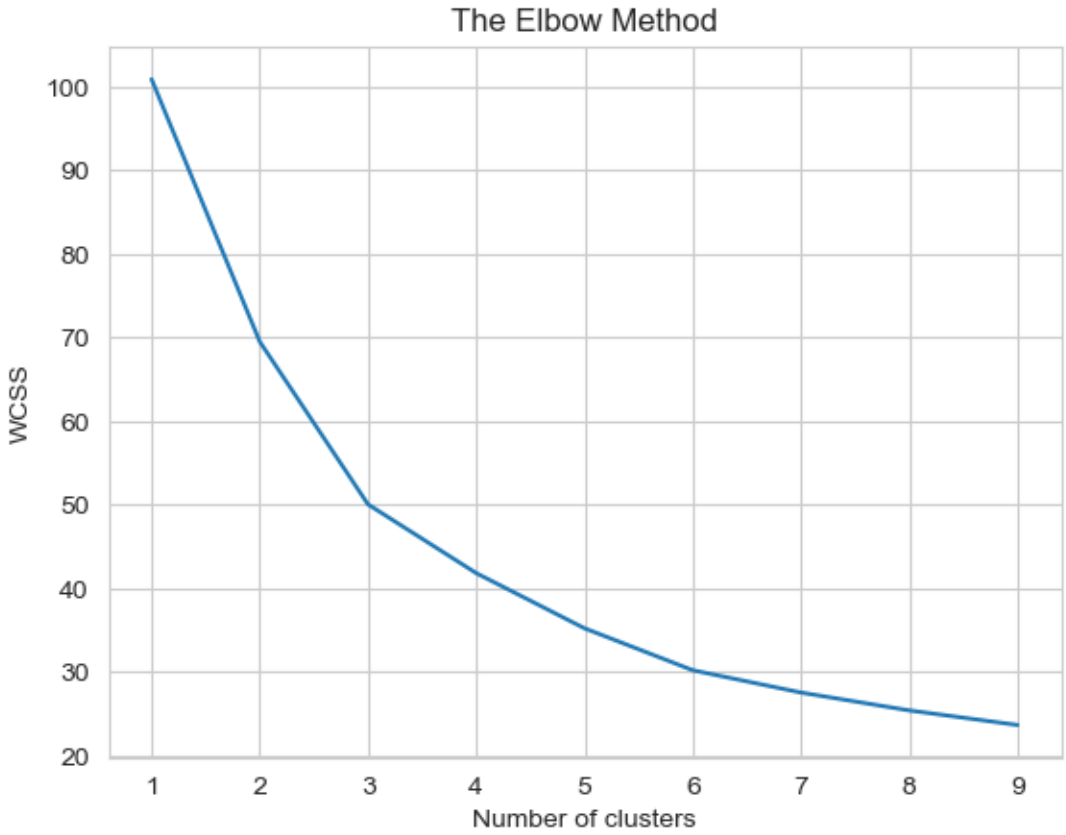
Note:

The Standard Scaler is a data preprocessing tool that standardizes the data by subtracting the mean and scaling to unit variance.

Apply Elbow Method

Note:

The "Elbow Method" is a technique to determine the optimal number of clusters in a k-means algorithm.



The Gain from this is the best number of clusters is 3.

Evaluated the K-Means cluster using Silhouette metrics.

The average silhouette score is: 0.292032236628384

Summary
Based on the average silhouette score of 0.292 suggests that the clusters are reas onably well-defined, but there may be some overlap between clusters.

Conclusion

- The variables we have considered in our analysis (such as economy, family, health, etc.) are not the only factors that influence happiness.
- There may be other variables or factors that are important for people's hap piness, but we have not considered them in our analysis.
- Cultural and social factors, environmental conditions, and personal circums tances are some of the factors that may affect happiness, but they are not i ncluded in the dataset or the variables we analyzed.

References:

- Lauren Erdelyi. (2020, MARCH). The Five Stages of The Data Analysis Process https://www.lighthouselabs.ca/en/blog/the-five-stages-of-data-analysis
- DataCamp. (2022, Dec). A Beginner's Guide to Predictive Analytics. https://www.datacamp.com/blog/predictive-analytics-guide
- SPSS. (2022). Shapiro-Wilk Test – Quick Tutorial. https://www.spss-tutorials.com/spss-shapiro-wilk-test-for-normality/

- Analytics vidhya. (2021,NOV). Understanding K-means Clustering in Machine Learning https://www.analyticsvidhya.com/blog/2021/11/understanding-k-means-clustering-in-machine-learningwith-examples/

- Rohit Sharma. (2022, AUG). Clustering vs Classification: Difference Between Clustering & Classification https://www.upgrad.com/blog/clustering-vs-classification/

Question 4:

Differentiate between Overfitting and Underfitting in Machine learning.

- Overfitting occurs when a machine learning model learns the training data too well, including the noise and outliers. As a result, the model performs poorly on new data that it has not seen before.
- Underfitting occurs when a machine learning model does not learn the training data well enough. As a result, the model performs poorly on both the training data and new data.

Exploratory Data Analysis (EDA) is a statistical method for inspecting a dataset to discover patterns, to summarize main features of the data, and to discover underlying relationships.

The objective of EDA is to understand the data and to make it easier to visualize and interpret.

The main steps to perform EDA are as follows:

- Data cleaning. This step involves removing errors, outliers, and missing values from the data.
- Data exploration. This step involves visualizing the data and using statistical methods to describe the data.
- Data reduction. This step involves reducing the dimensionality of the data or identifying clusters of similar data points.
- Data visualization. This step involves creating plots and charts to help visualize the data.
- Data interpretation. This step involves interpreting the results of the EDA and drawing conclusions about the data.