# A CAREFUL EXPLORATORY DATA ANALYSIS OF "WORLD HAPPINESS REPORT OF 2023"

*Author: José Bento Ferreira Montenegro*

*Jul/02/2024*

*[bentomontenegro@gmail.com](mailto:bentomontenegro@gmail.com)*

# 1. Introduction

Exploratory Data Analysis (EDA) is a fundamental step of machine learning workflows. It is a powerful toolbox that allows us to efficiently extract statistical information about the data even before making a model, as well as to select which attributes as characteristics of the data are essential for the modeling process.

In the present work we analyze the data set “[World Happiness Report of 2023](https://www.kaggle.com/datasets/joebeachcapital/world-happiness-report-2013-2023/discussion/433225)”, conducting all the best-practices steps of EDA, as well as a formal hypothesis test to exemplify the major role of statistical analysis in today’s industrial scenario. More specifically, it is well known that raw data sets almost never come ideally suited for modeling purposes. We then show through code snippets and figures how the process of cleaning, removing and renaming columns, re-scaling attributes, handling outliers, analyze correlations and testing hypotheses can be implemented on real data science workflows.

This is actually my final project for the EDA module of the [IBM Machine Learning Professional Certificate](https://www.coursera.org/professional-certificates/ibm-machine-learning) course on [Coursera](https://www.coursera.org/)’s platform. I hope the reader will enjoy it.

# 2. Brief description of the data

As said above, we shall analyze the data set "World Happiness Report of 2023". This data set contains information about the happiness scores reported from population samples of several countries around the world, as well as economic and political factors specific of each country. Of course, taking “Hapiness score” as target and such factors as features, the main purpose of the data set is to understand what exactly might influence individual reports at each country.

Let us have a look at the initial attributes of the data set.

**import pandas as pd**

**filepath = 'data/World Happiness Report 2023.csv'**

**data = pd.read\_csv(filepath)**

**data.info()**

Output:

# *Fig. 1: Attributes of the data set*

# 3. Initial plan for data exploration

As suggested by the list of attributes above, previous data analysis has been set forth on this data set. In order to make it more suited for the purposes of the present work, I suggest the following steps, which will be explored in the sections to come:

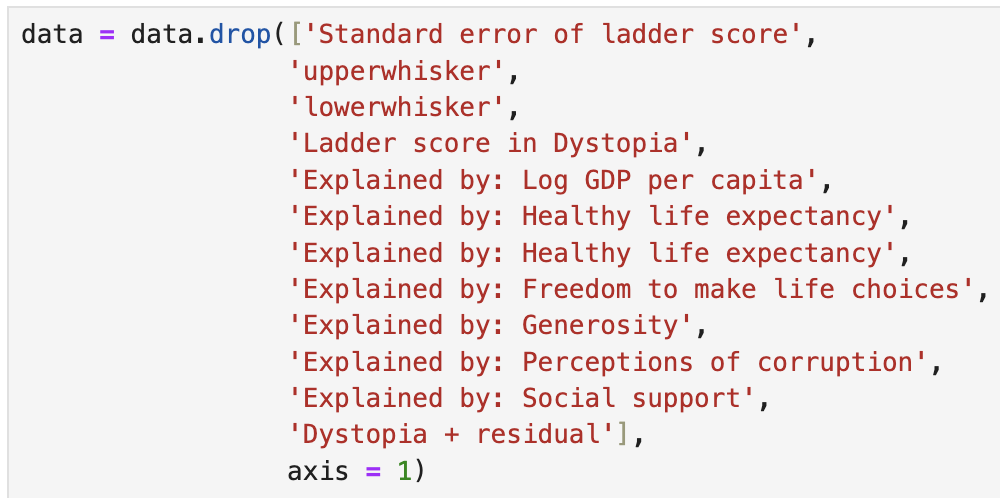
1. Remove columns that refer to the previous data analysis
2. Rename certain columns more concisely
3. Identify null values and decide what to do with them
4. Re-scale attributes, if necessary
5. Identify outliers and decide what to do with them
6. Plot correlations between features and target
7. Extract information from the correlation plot
8. Propose three relevant statistical tests for this data
9. Explore one of such tests using formal statistical analysis

# 4. Data cleaning and feature engineering

**4.1 Removing columns**

First of all, we would like to conduct a brand new EDA (Exploratory Data Analysis) on this data. To this end, it is useful to begin by removing all columns that make reference to any previous analysis conducted on it. Such columns are:

1. Upperwhisker
2. Lowerwhisker
3. Standard error of ladder score
4. Ladder score in Dystopia
5. All the “Explained by” columns
6. Dystopia + residuals

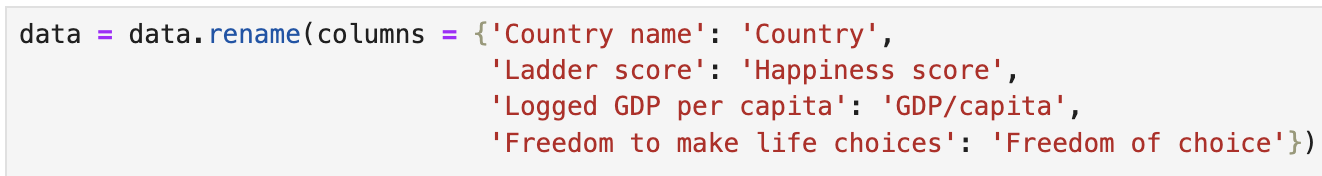
We remove them at once using data.drop method.

# *Fig. 2: Dropping columns that are irrelevant for our analysis*

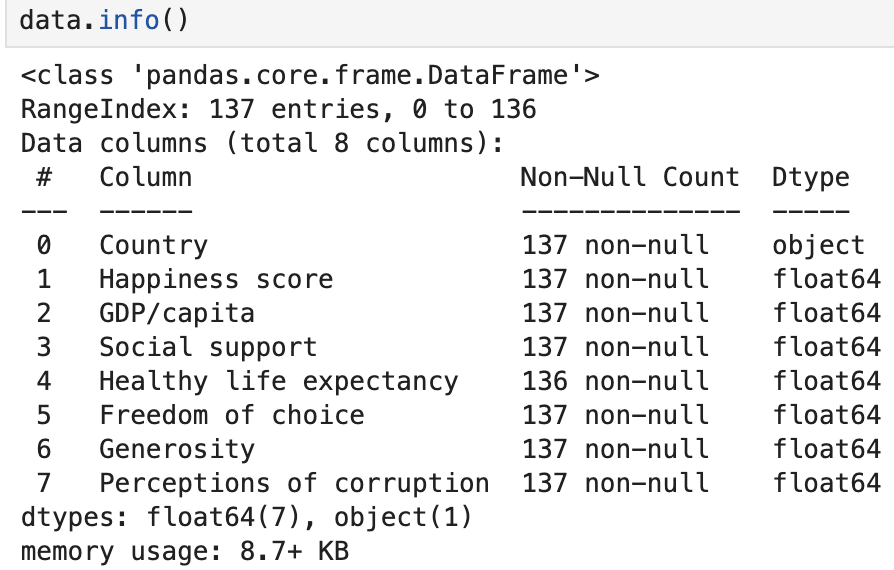
**4.2 Renaming columns**

Going further, let us use data.rename method to make the following change of names:

1. Ladder score → Happiness score
2. Logged GDP per capita → GDP/capita
3. Freedom to make life choices → Freedom of choice



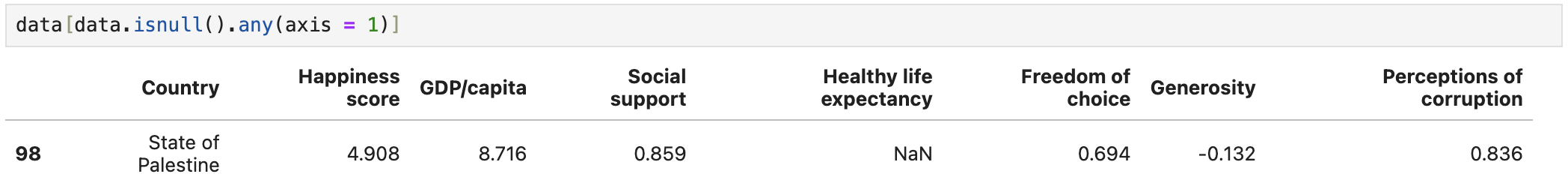
# *Fig. 3: Renaming columns*

Now our list of attributes looks much cleaner:

# *Fig. 4: Data attributes after cleaning and renaming*

**4.3 Null values, re-scaling and outliers**

**4.3.1 Null values**

Now, as suggested by the above figure, there is only one missing value that we have to deal with. It is relative to the feature “Healthy life expectancy” for the country “State of Palestine”, as can be confirmed below.

# *Fig. 5: The only row containing a null value*

We could simply remove it in future calculations involving this specific attribute, but let us replace it by the mean value of the whole attribute “Healthy life expectancy” for simplicity. This might not be the most accurate approach to, but hopefully this replacement will not compromise future models significantly. If it does, for instance, we can simply come back, remove it and check if this results in a better model.

Let us use data.fillna to implement the replacement:

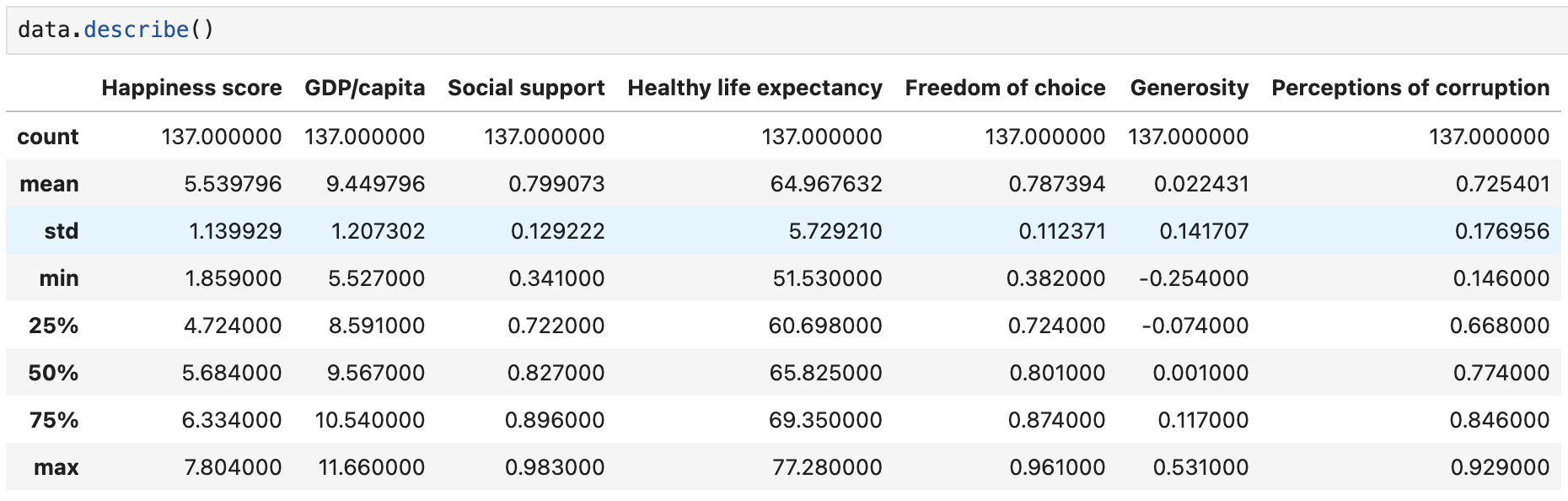


# *Fig. 6: Row “State of Palestine” after replacing the null value in “Healthy life expectancy"*

**4.3.2 Re-scaling**

Now we check if any attribute of the data needs re-scaling, as making sure that all attributes have similar scales is a good practice for regression modeling.

We start by looking at the statistical parameters of the data:



*Fig. 7: Statistical parameters of the data*

We can see that the mean values of “Social support”, “Healthy life expectancy”, “Freedom of choice”, “Generosity" and “Perceptions of corruption” are not in the same scale as the target variable “Happiness score”.

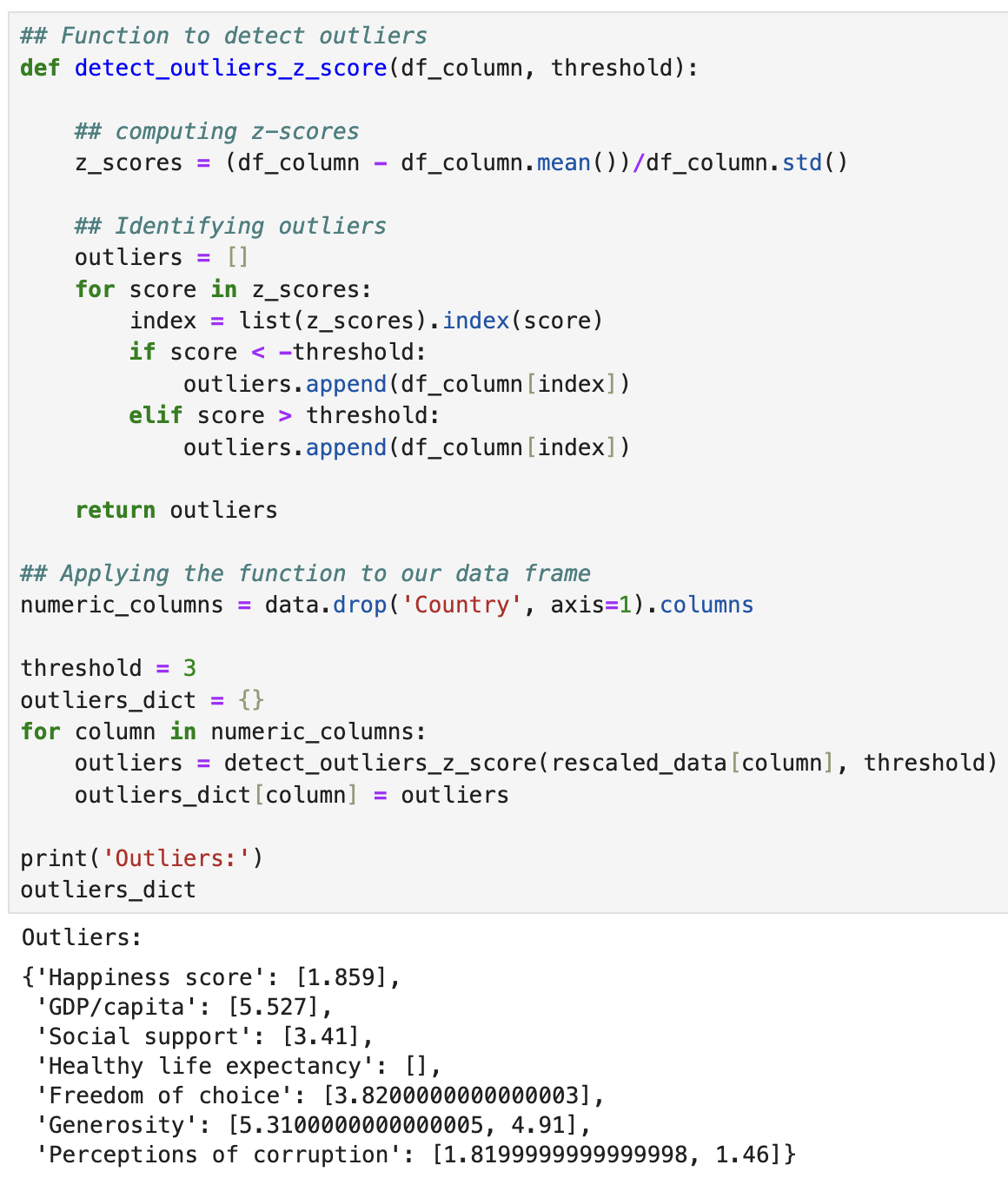
The scales are not drastically far from each other, and we could use several methods to fix this, such as z-score normalization and min-max scaling. In general, the former can make the data particularly suited for linear regression models, and the later has the advantage of being very sensible to outliers. However, I propose to keep things straight here and simply multiply each column by an appropriate power of 10, so that all attributes lie in a range similar to the range of the target, which [0, 10]. This can preserve most of the data structure and distribution for now, allowing us the possibility to choose any other scaling when modeling. See Fig. 8.

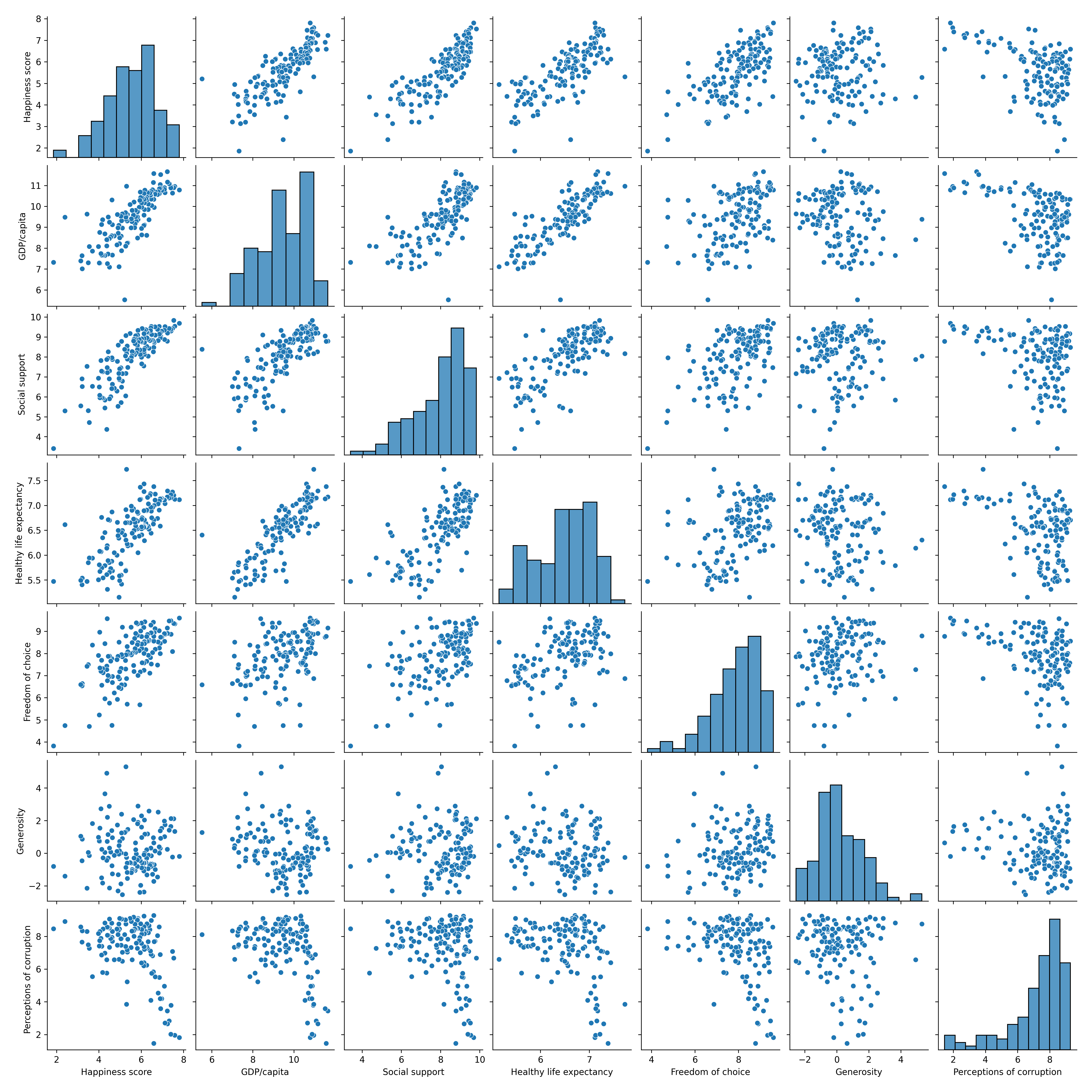
**4.3.3 Outliers**

Now that we re-scaled our data, it is a good practice to look for outliers that may interfere in any future analysis. As always, there are several methods available for this, such as box plots and z-score analysis. In this work I will choose the later.

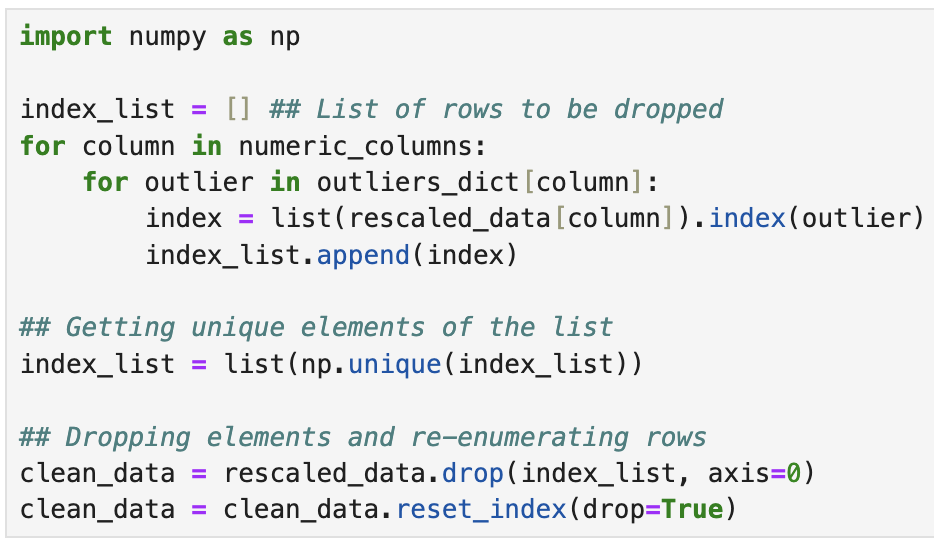
In summary, z-score analysis consists in calculating the z-score ((x - xmean)/σ) for every point x of a given attribute with mean xmean and standard deviation σ, and comparing it to a certain *threshold value* T. More specially, the z-score measures how far x is from xmean in units of σ. If x has a z score that is less than -T or greater than T, it is considered an outlier. It is a common practice in statistics to choose T = 3.

*Fig. 8: Statistical parameters after re-scaling*

In Fig. 9 we create a function that identifies the outliers of a given column. We then apply this function to our re-scaled data set and drop all the rows containing outliers, as shown in Fig. 11.****



*Fig. 9: Using z-value scores to identify outliers*

****

*Fig. 10: Removing all rows containing outliers*

*Fig. 11: Pairplot correlations between data attributes*

**4.5 Correlations**

Now that we organized, cleaned and re-scaled our data, it is useful to check how its attributes correlate among themselves. This is a valuable step in modeling, where we have to decide which features to use and which features are really independent from each other.

A simple way to start is to use seaborn’s paiplot function, as shown in Fig. 11. From there we can already infer many interesting conclusions, for example:

* The GDP/capita of individuals seems to grow when they are happier, have high healthy life expectancy and social support.
* The GDP/capita does not seem significantly correlated with how much freedom of choice individuals report to have, nor with their generosity and perception of corruption. This is due to the fact that the data points are mostly disperse in such plots and do not seem to follow any regular pattern.

Since we took “Happiness score” as our target variable, it also valuable to write down some conclusions relative to it explicitly.

* It seems to grow almost linearly with "GDP/capita", "Social Support", "Healthy life expectancy" and "Freedom of choice”, suggesting that a linear regression model may be a good fit for prediction and interpretation.
* It seems to have no regular alongside “Generosity”, suggesting that individuals not necessarily feel happier as they feel more generous.
* Although at first it seems to decrease with “Perception of corruption", there is a certain point where this behavior simply diffuses. This suggests that a linear regression model might be a good fit only when we deal with populations that perceive less corruption around them.

# 6. Hypothesis testing

This is a rich data set and many hypothesis can be formulated for it. Three examples are:

1. On average, Asians are as happy as Europeans
2. The GDP/capita has a very significant influence on Happiness scores
3. The variance of samples of Middle-East is higher than the variance of samples of Asia

We can always use formal statistical analysis to test such hypothesis. For example, we could use a t-test to test the first one, a z-test (or even a regression model) to test the second, and a f-test to test the third. Unfortunately, explaining all these procedures is out of the scope of the present text, and we encourage the reader to find more about them online.

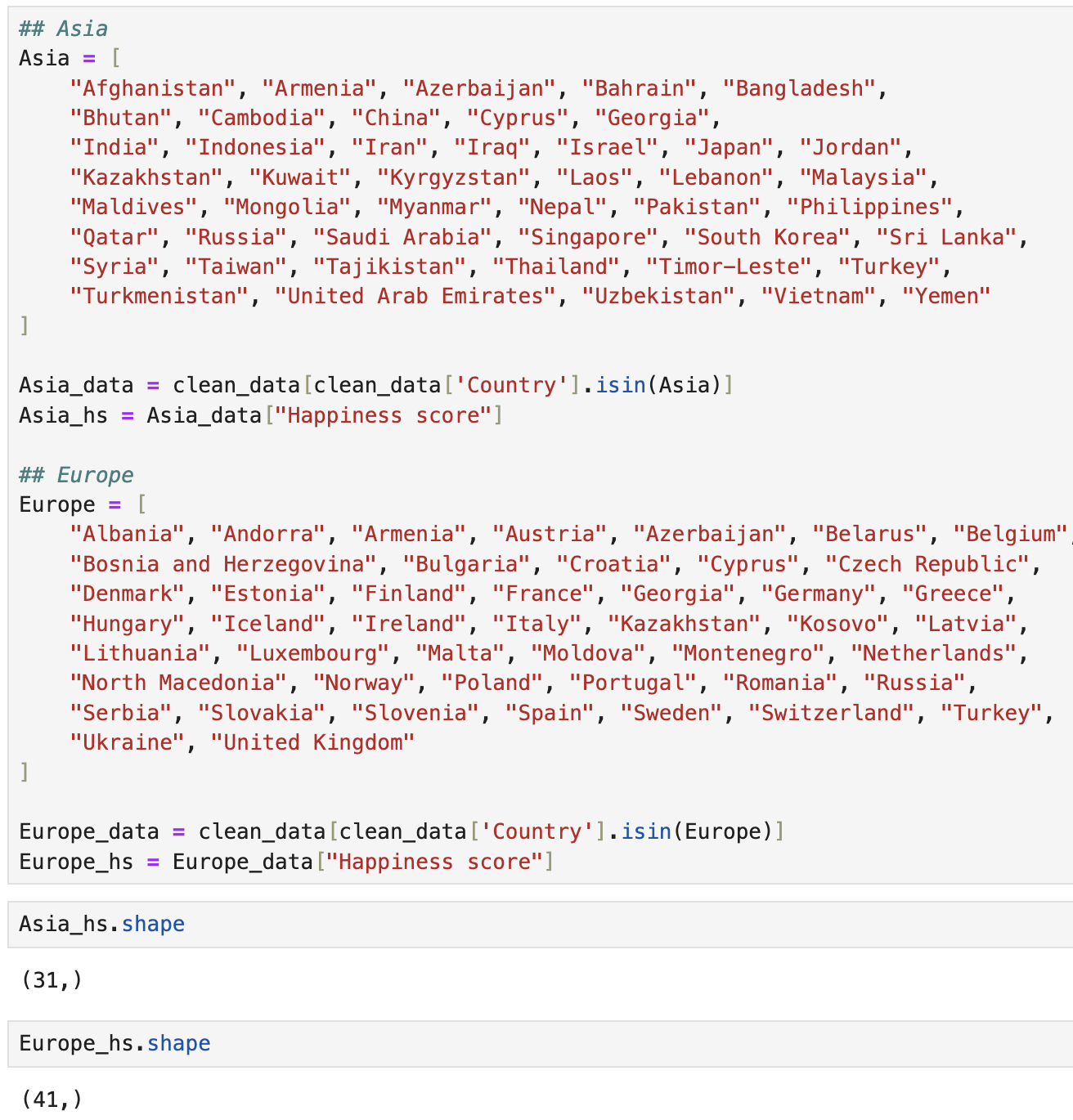
As an illustration, let us test the first hypothesis. The null and the alternative hypotheses can be formulates as:

**H0:** μEurope - μAsia ≠ 0 (*The Happiness scores of Asia and Europe are significantly different*)

**H1:** μEurope - μAsia = 0 (*The Happiness scores of Asia and Europe are not significantly different*)

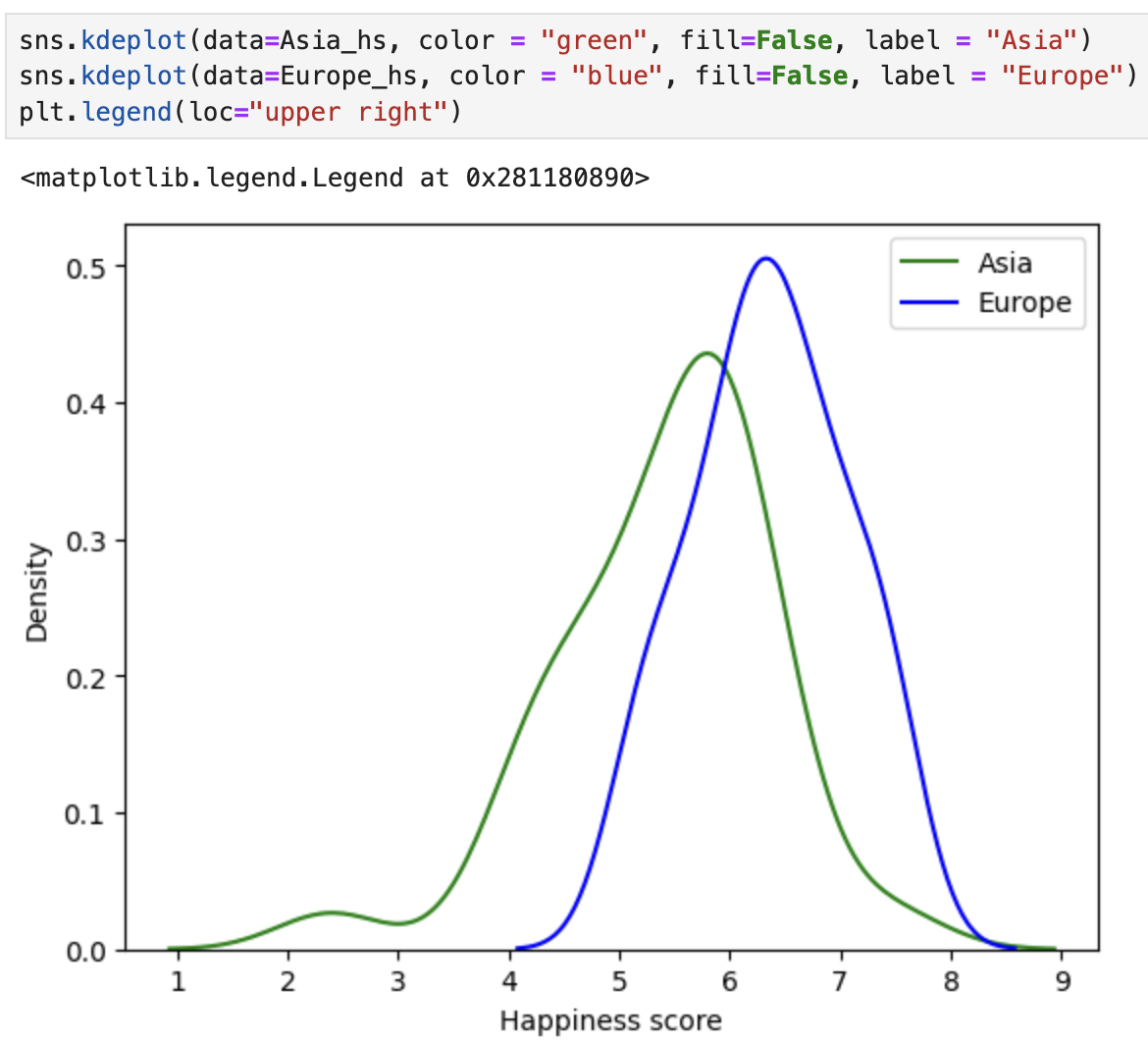
Now, let us fix the significance level α = 0.05 (5%), as usual. If the p-value of the test is less than this value, we conclude that the two means are, in fact, significantly different. Thus, we cannot reject the null hypothesis.

Let us first sample the countries of Asia and Europe into two separate data frames:



# *Fig. 12: Sampling countries from Asia and Europe*

We can see that we have information about 41 countries of Europe and 31 of Asia. However, we know that Europe and Asia have approximately 50 and 48 countries, respectively (“approximately” because it depends on the geopolitical definitions used). Therefore, we must then keep in mind that we might be dealing with somewhat restricted samples, and this analysis can be significantly refined if more data is available

In Fig. 13 we can have a look at the target distribution of the two regions. It is clear (at least visually) that the two sets of Happiness scores follow significantly different Normal distributions, but that their means (midpoint of each curve) are actually not so distant. (Nonetheless, we cannot conclude that they are significantly different without a formal hypothesis test because statistical analysis of samples always involve a significant degree of uncertainty in estimating statistical parameters such as mean values.)

# *Fig. 13: Happiness score distributions of Asia and Europe*

Now we use scipy’s stats function to conduct the hypothesis testing more efficiently.

# *Fig. 11: Testing the hypothesis that Asia and Europe have equal average happiness scores.*

**Conclusion:** As the p-value of the test is very low, we cannot reject the null hypothesis that the two mean Happiness scores are significantly different. In particular, from our samples, it is suggested that Happiness scores of Europe are (on average) greater than Happines scores of Asia.

However, it should be stressed again that we are dealing with a somewhat restricted sample of Asian countries. This suggests that more data is highly desirable to support the result of the test.

# 7. Conclusion

In the present work, we performed an exploratory data analysis on the data set “World Happiness Report of 2023”. We followed the well-known best-practices of data cleaning and feature engineering, namely: dealing with null values, renaming the data attributes more concisely, re-scaling data attributes for future modeling, looking for correlations between data attributes, interpreting these correlations, and conducting a formal hypothesis test to exemplify how statistical analysis can be a powerful tool to extract predictions from the data before modeling.

From our analysis we could determined insightful characteristics about the happiness scores reported from countries around the world. For example, we saw that happiness scores seem to be significantly correlated with the net GDP/capita and level of Social Support of each country. We also found evidence, via a formal hypothesis test, that Asia and Europe have, on average, significantly different happiness scores, though more data about Asian countries is highly desirable to support this conclusion.