Forecasting Quality of Life: The Future of Nations

Data Science Lab Project

Cervini Stella
Simone Mattia
Montalbano Daniel
Sabino Giuseppe

Contents

1 Introduction

In an era characterized by unprecedented data availability and globalization, the quest to objectively measure and estimate the quality of life across the globe has become increasingly important. This is why we have chosen to address this topic in our Data Science Lab project.

The heart of this work is the "**Life Index**", a scalar metric that ranges from 0 to 1, which serves as a compass to measure and delineate the quality of life within a given country. The development of this Index is based on several data sources: political (stability and freedom), economic (GDP) and demographic (population, birth rates and life expectancy).

The essence of this project is not only to assess the present, but also to look into the future: using historical data from the past 15 years, our research attempts to predict the evolution of these crucial indicators, including the synthetic Life Index. Using time series analysis techniques, we aspire to offer insights into how these indicators will develop, thus providing a look into the future direction of nations around the world.

Moreover, this research extends beyond the borders of individual nations: using clustering methodologies, we tried to uncover patterns and similarities between nations based on these key indicators. In addition, we conducted data aggregations on a continental level, revealing regional trends, disparities and opportunities.

This project provides practical applications that have the potential to have an impact on people's lives. It is possible to image an online service, supported by a sustainable business model, that allows users to identify countries that match their aspirations, both current and future, whether for permanent residence, travel or investment.

2 Dataset

2.1 Data sources

2.1.1 Freedom House organization

The input source for the project is provided by the data offered by the **Freedom House Organization** [1]: a non-profit organization that promotes democracy and human rights around the world. The dataset [2] consists in scores and a status regarding political rights and civil liberties for each country and a select group of territories, from 2003 to 2023.

A *total index* has been defined as a value, from 0 to 100 points, that is assigned to each country based on 10 political rights indicators and 15 civil liberties indicators, which take the form of questions. Scores regarding the political rights are summarized by the *PR Index*, which ranges from a value of 0 to 40, and the 10 questions are grouped into three subcategories:

- A: Electoral process;
- B: Political pluralism and participation;

- C: Functioning of government.

The civil liberties are summarized by the *CL Index*, which ranges from a value of 0 to 60, and the 15 questions are grouped into four subcategories:

- D: Freedom of expression and belief;
- E: Associational and organizational rights;
- F: Rule of law;
- G: Personal autonomy and individual rights.

An important indicator is the *Status*, it can assume the values: Free, Partly Free, Not Free Status. The status is determined by the combination of the overall score awarded for political rights and the overall score awarded for civil liberties, after being equally weighted, as shown in the figure 1.

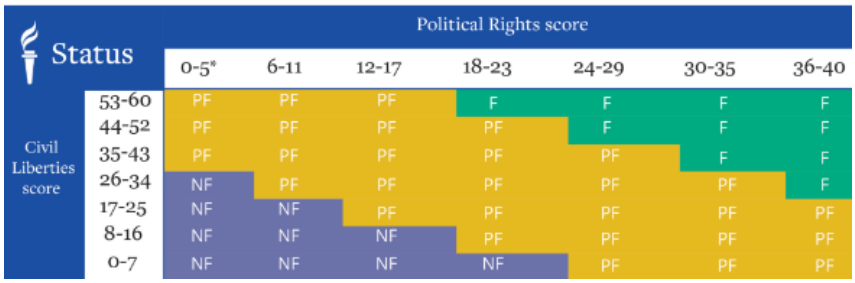

Figure 1: Status indicator

For the purposes of this project, we used only the *Status* and *Total* indices.

2.1.2 The world bank

The main source of data is provided by the **World Bank Organization** [3]: a global institution that provides financial and technical assistance to developing countries. The organization maintains a platform [4] that provides free and open access to data on development.

Specifically, the following indices were used:

- Population [5]: number of people living in a particular area;
- GDP per capita [6]: average income of a country's citizens in dollar;
- Political stability and absence of violence/terrorism [7];
- Birth rate [8]: number of births per 1,000 people in a year;
- Life expentancy [9]: average number of years a person is expected to live.

2.2 Data Integration

The integration of the different data sources was carried out using the country name classification used by the first source (section 2.1.1): this required mapping between different names or the handling of special cases (e.g. territories not present in both data sources). A practical example is related to the territory of Palestine, which in the first source is divided between Gaza Strip and West Bank: in the final dataset we have kept this division, keeping the classification of the first source as standard, and we have assigned the same index values of the second source to both terristories.

In order to manage missing data, it was decided to remove rows with too many missing fields. Some indicators from the second source did not have the same completeness as the first, which led to the reduction of the historical period from 2006 to 2021. So, the following countries have missing data for the 15-year historical period chosen for the analysis: Djibouti (only 9 years), Gaza Strip (only 11 years), Kosovo (only 14 years), Puerto Rico (only 11 years), Somalia (only 9 years), South Sudan (only 4 years), Turkmenistan (only 14 years) and West Bank (only 11 years).

In conclusion, the dataset is composed by the follwing fields:

- Country/Territory: name of the country/territory;
- Region: geographical area;
- Edition: year to which the data refers;
- Country code: three-digit acronym used to identify a country;
- Status and Total (defined in section 2.1.1);
- Population, Political stability, GDP, Birth rate and Life expectancy (the latter defined in section 2.1.2).

3 Life Index: a new index for quality of life

The **Life Index** represents a composite index that aims to measure the quality of life or well-being of a specific geographical area or group of countries. This index is calculated by considering a set of socio-economic and demographic variables, which are normalized and then combined to obtain an overall value that reflects a comprehensive assessment of the situation in a given country.

The Life Index is obtained through the arithmetic mean of the following five normalized variables: *GDP*, *Total*, *Political Stability*, *Birth Rate* and *Life Expectancy*. This index reflect an overall assessment of quality of life or well-being based on the five dimensions considered. A higher score would indicate better quality of life or higher well-being, while a lower score would indicate the opposite.

Figure 2 reveals a distinct pattern regarding the quality of life across continents. Both America and Europe exhibit notably higher life indices, suggesting better overall living conditions for their populations. In contrast, Asia and Africa are areas of particular concern when it comes to life expectancy and overall quality of life. These continents appear to have lower life indices, indicating challenges related to healthcare, socio-economic development, and various factors affecting the life of their inhabitants. The data displayed underscores the significant disparities in living conditions and life prospects between these continents, highlighting the need for targeted efforts and interventions to improve the quality of life.

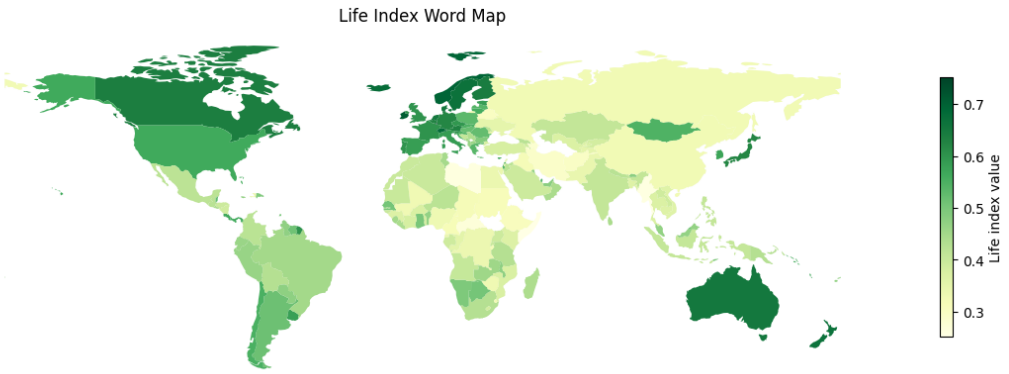

Figure 2: Life index world map

4 Exploratory analysis

An exploratory analysis of the data was conducted for understanding data and finding any relevant pattern.

4.1 Descriptive statistics

//	Total	Population	GDP	Political Stability	Birth rate	Life expectancy
mean	0.615341	3.860036e+07	18476.663436	-0.069511	21.639045	70.795916
std	0.281556	1.429661e+08	20384.519031	0.948938	10.618244	8.297434
min	0.010000	1.003000e+04	528.871700	-2.826402	5.100000	42.914000
25%	0.370000	2.012647e+06	4073.926085	-0.670838	12.000000	64.815000
50%	0.640000	7.583269e+06	11435.268434	0.010891	19.334000	72.457220
75%	0.890000	2.702694e+07	26625.061810	0.741739	29.686000	77.038000
max	1.000000	1.412360e+09	163219.491990	1.639301	50.096000	84.560000

Table 1: Relevant variable's summary (referring to 2021).

From table 1 displays it is possible to retrieve some key observations. First of all, there is a **wide range** of values for each variable. For example, the population size ranges from 10,030 people to 1,412,360,000 people while the birth rate ranges from 5.1 births per 1000 people to 50.09 births per 1000 people. The **standard deviation** for each variable is also relatively large, which indicates that there is a lot of variation in the data. Wide ranges and

high standard deviations suggest that there are quite large disparities between the countries under study.

Regarding the **distributions**, we can state the following:

- The population size is right-skewed, which means that there are a few countries with very large populations and many countries with smaller populations.
- The GDP is also right-skewed, but to a lesser extent than the population size. This means that there is a greater degree of equality in the distribution of GDP than in the distribution of population size.
- The political stability score is approximately normally distributed, indicating a fairly even distribution of political stability scores across the countries.
- The birth rate is left-skewed, meaning that there are a few countries with very high birth rates and many countries with lower birth rates.
- The life expectancy is approximately normally distributed, suggesting that there is a fairly even distribution of life expectancies across the countries and then worldwide, even if the difference between the minimum and maximum value registered is quite large (approximately 40 years).

These observations suggest that there are a number of factors that contribute to the variation in the values of the variables. For example, the population size is likely influenced by factors such as geography, natural resources, and economic development, while the GDP is reasonably influenced by factors such as the size of the economy, the level of industrialization, and the trade balance. Regarding the political stability score, can could infer that the variable is linked to other aspects of a country's life, such as the level of corruption and the level of violence. In particular, these latest components are quite difficult to measure. The birth rate could depend on birth control diffusion and overall cultural norms, while life expectancy is heavily dependent on the quality of healthcare and the level of sanitation.

Finally, Life Index is simply a combination of all of these factors, resulting in an effective synthesis of many aspects that characterize the intern situation of a country.

4.2 Relationships between variables

Further analysis of the data could be used to investigate the relationships between these variables and to identify the factors that have the greatest impact on the well-being of the population. For this reason, in Figure 3 is reported the **correlation matrix** between all variables. Here Life Index is not considered, since it is supposed to be correlated with all the variables, considering that it is composed by their normalized values.

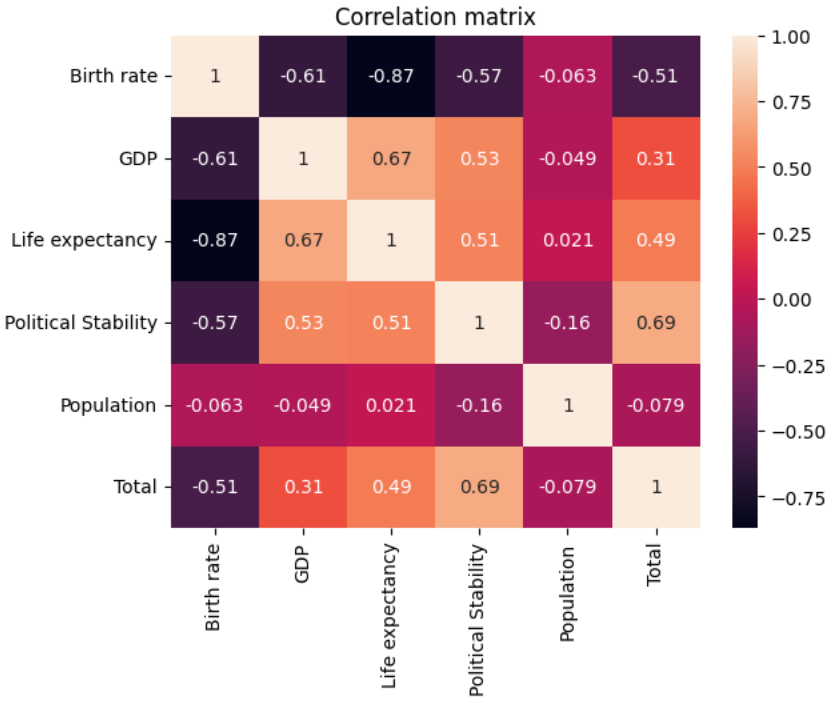

Figure 3: Correlations.

A **correlation coefficient** is a measure of the linear relationship between two variables. A correlation coefficient can range from -1 to 1. A positive correlation coefficient indicates that the two variables are positively correlated, meaning that they tend to move in the same direction. Moreover, a negative correlation coefficient indicates that the two variables are negatively correlated, meaning that they tend to move in opposite directions. A correlation coefficient of 0 indicates that there is no linear relationship between the two variables.

The correlation matrix shows that there are a number of **strong correlations between the variables**. The strongest (negative) correlation is between birth rate and the life expectancy (-0.870), this means that countries with higher birth rates tend to have lower life expectancies. There is also a strong correlation between the birth rate and the GDP (-0.611), this means that countries with higher birth rates tend to have lower GDPs. Regarding the positive correlatoins, the strongest of them is observed between the GDP and the life expectancy (0.674), indicating that countries with higher GDPs tend to have longer life expectancies. The political stability and the population are also positively correlated (0.528): countries with higher political stability tend to have larger populations.

The correlation matrix can be used to identify potential relationships between the variables. However, it is important to note that **correlation does not equal causation**. Just because two variables are correlated does not mean that one causes the other. There may be other factors that explain.

4.3 Life Index

The figure 4 show that the Life Index, in the majority of regions, has fallen in recent years, despite having remained constant or increasing in previous years.

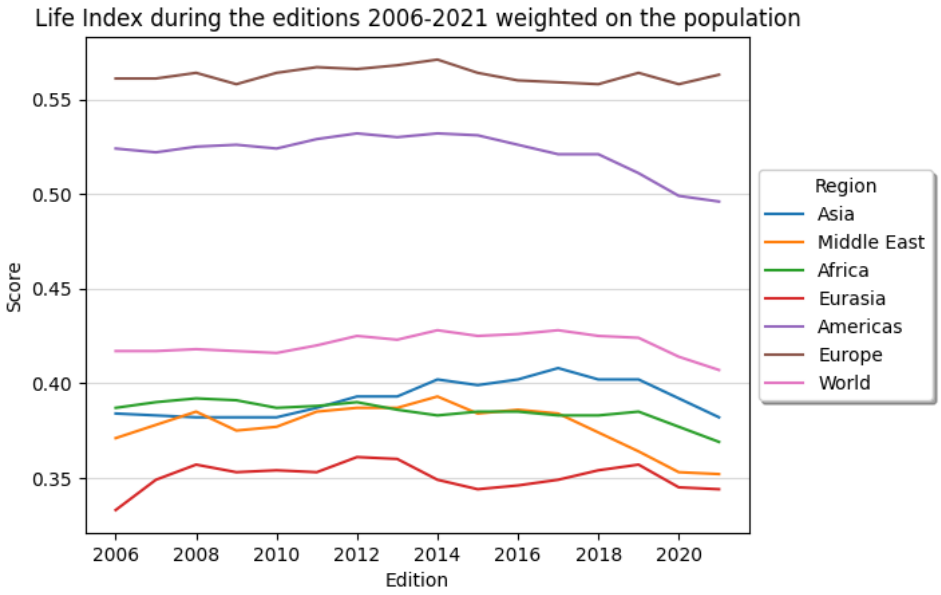

Figure 4: Life Index trend from 2006 to 2021 weighted on the population.

Europe has the lead, recording the highest Life Index, followed by *Americas*. These two regions have a large detachment from all other regions and even the global Life Index, indicating very high overall life conditions. *Eurasia* has the lowest Life Index, followed by the *Middle East* and *Africa*.

The wide range of values in the Life Index metric could be mainly attributed to **differences in living standards, healthcare, and education**, which contribute to high Life Index for some regions, unlike other which are more penalized. The **political stability** of these regions has also played a role. The **high poverty rates** and political instability in many African countries have hindered their economic development and led to a decline in the Life Index, while the ongoing conflict in the Middle East has also had a negative impact on the Life Index in this region.

Figure 5 above shows the Life Index for each region obtained in the various editions of the FIW Survey.

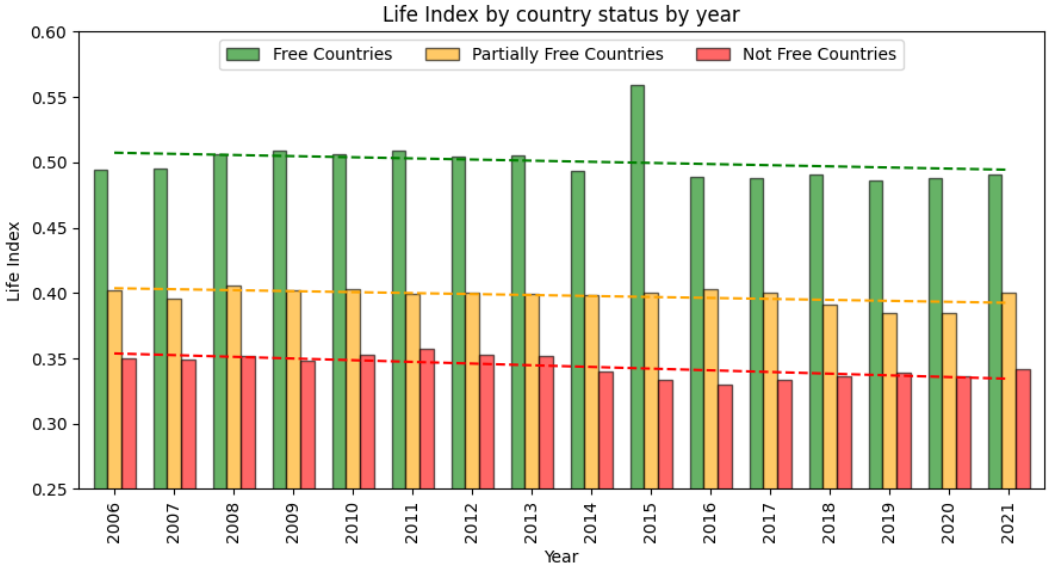

Figure 5: Life Index by country status (F, PF, NF).

From the image, some **trends** that can be observed:

- The **average Life Index for Free countries** has been fluctuating over time. In 2006, the average Life Index for Free countries was 0.494 and, by 2021, it had decreased to 0.491. Despite this slightly decrease, the value from Life Index suggests an overall well-being of the population in Free countries.
- The **average Life Index for Partially Free countries** has been relatively stable over the past years. In 2006, the metric had a value of 0.402, while in the last year recorded it is equal to 0.400. This indicates that the general welfare has not changed significantly over time.
- The **average Life Index for Not Free countries** has been decreasing over time, from 0.350 in 2006 to 0.342 in 2021. The prosperity of the population in the countries included in this category is in general worsening.

It is important to note that these are just trends, and there are always exceptions. There are some Free countries where the Life Index has decreased, and there are some Not Free countries where the Life Index has increased. Similar comments can be made on the different years under study, since there are periods with an increasing trend and other with a decreasing tendency. However, the overall trends suggest that the Free countries are doing a better job in keeping high the well-being of their populations than the Not Free countries.

There could be multiple factors that influence these movements. Some possible explanations for the phenomena are reported below:

- Free countries tend to have more stable political systems and stronger rule of law, which can create a more favorable environment for economic growth and development.

They also tend to have more open societies, which can lead to greater social mobility and opportunities for people to improve their lives.

- Not Free countries, on the other hand, often have authoritarian governments that suppress dissent and limit economic opportunities. This can lead to lower standards of living and a decline in the overall well-being of the population.

It is important to continue to monitor these trends and to identify the factors that are contributing to them.

5 Cluster analysis with the K-means algorithm

5.1 K-Means algorithm

The **K-means** algorithm is an unsupervised machine learning algorithm. Its primary function is to attempt to identify distinct groups within a dataset, where the elements of each group are as similar to each other as possible and as different as possible from the elements of other groups.

This algorithm begins by randomly selecting K distinct points on the plane, called **centroids**. Subsequently, it cyclically performs the following two steps:

- *Assignment of Points to Centroids*: for each data point in the dataset, it determines the nearest centroid by calculating the Euclidean distance between the point and the centroids. The point is then assigned to the nearest centroid, thus forming a cluster. In other words, each point is associated with the cluster whose centroid is closest in terms of Euclidean distance. Considering x_i as the point of the dataset and C_j as the centroid, the Euclidean distance is defined as following:

$$\text{Euclidean distance}(x_i, C_j) = \sqrt{\sum_{k=1}^{n} (x_{i_k} - C_{j_k})^2}$$

- *Recomputation of Centroids*: it calculates the new centroid for each cluster. This new centroid is the average of the points belonging to that cluster. In practice, it means that the centroid of a cluster moves towards the average position of the points within that cluster. If C_i is the centroid of cluster i, the formula for calculating the new centroid is:

$$C_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i_j}$$

After calculating the new centroids for all clusters, all data points are reassigned to the new centroids. This process is cyclically repeated until there are no further changes in the assignment of points to clusters, indicating convergence or a maximum predefined number of iterations have been executed.

5.2 Application in this context

In this project, we will employ the K-means algorithm to explore a rich dataset of socio-economic and demographic data originating from various regions or countries. Our aim is to identify clusters of similar entities, unveil which variability among the features is most significant, and uncover trends that might go unnoticed in a traditional analysis.

The main purpose is to **identificate homogeneous groups**. In fact, clustering can assist in identifying groups of countries that share similar characteristics. For instance, you can identify clusters of countries with similar per capita income, life expectancy, birth rates, and so on. This can be valuable in gaining a better understanding of regional socio-economic dynamics.

The first step in considering all available observations is to calculate the average of values over the years, in order to incorporate it into the development of clusters. Furthermore, to achieve a more effective segmentation, only the more relevant columns have been selected (*GDP, Total, Political Stability, Birth Rate, Life Expentancy*). All these values have been normalized to ensure a more meaningful comparison between the various columns, that is **min-max normalization** has been employed, which allows for normalized values to range between 0 and 1.

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

5.3 Results

At this point, one of the most important steps was to determine the ideal number of clusters to use. To do this, the **elbow method** was employed, in which the Silhouette value is used to identify the point at which increasing k will cause only a very small decrease in Silhouette, while decreasing k abruptly increases its value.

$$s(i) = \frac{\max\{a(i), b(i)\}}{b(i) - a(i)}$$

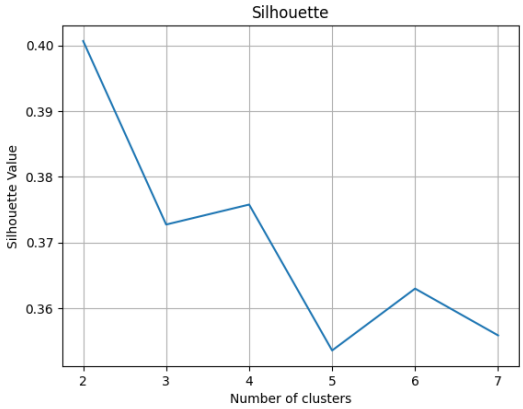

Figure 6: Elbow Method.

Considering the Figure 6, it can be observed that there is no actual *"elbow point"* as explained earlier. Therefore, it was decided to use the point that minimizes the silhouette value since this does not lead to overfitting issues. Thus, for the clustering process, **5 clusters** will be used.

At this point, the clustering operation has been carried out using the K-means algorithm, preceded by a **PCA (Principal Component Analysis)** operation to enable cluster visualization through a scatter plot providing the following visualization as output:

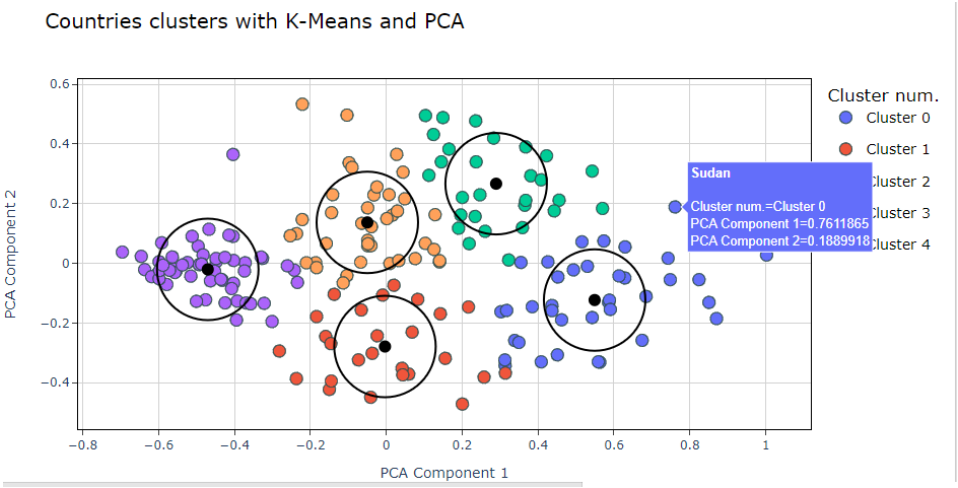

Figure 7: Graphical representation of the cluster with PCA.

From the graph in Figure 7, a **fairly clear division** between the 5 clusters can be observed, each marked by a different color. The black points represent the **centroids** of each cluster, and the highlighted area around them helps identify which observations are closer to each centroid. It is worth noting that the graph is *dynamic*, providing an immediate insight into the division of countries into each cluster. This scatterplot is useful primarily for highlighting the differences in distance between the various clusters, that is, how each cluster has been defined. However, it does not provide insights into the meaningful information for our specific case, as PCA collapses the values into just two points.

For this reason, a second visualization is provided in figure 8, namely a radar chart that displays the normalized values for each cluster. These charts are effective in showing how different variables contribute to an outcome. Each variable is represented by a spoke or line on the chart, and the area within the polygon formed by the lines provides an overall representation of the relationships between the variables. It is useful for visually **comparing multiple categories** and also **highlights various strengths and weaknesses**. As before, this chart is also dynamic, allowing you to observe one cluster at a time or compare two or more of them.

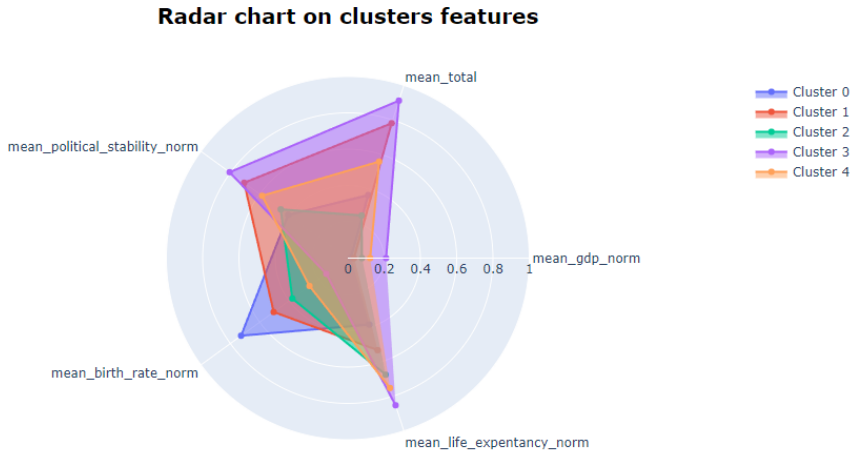

Figure 8: Radar chart on cluster features.

To provide a numerical and more immediate view, the table 2 is also provided. From here, the characteristics from the clusters are more easily comparable and comments on each cluster can be found below.

Cluster	Mean Birth Rate	Mean GDP	Mean Life Expectancy	Mean Political Stability	Mean Total
0	38.0	3132.0	59.0	0.406	0.365
1	28.0	6060.0	65.0	0.706	0.781
2	22.0	13219.0	71.0	0.457	0.247
3	12.0	34857.0	78.0	0.805	0.912
4	17.0	20540.0	74.0	0.584	0.559

Table 2: Clusters summary.

- **Cluster 0**: this cluster represents countries with high birth rates, low per capita income, relatively low life expectancy and moderate political stability. These countries may face some socio-economic challenges.
- **Cluster 1**: the cluster contains countries with lower birth rates compared to the first cluster and medium to high per capita income. On the other hand, these countries have decent life expectancy and strong political stability. For these reasons, we can state that they are in a relatively favorable position.
- **Cluster 2**: this cluster portrays territories with low birth rates, high per capita income, and high life expectancy, while political stability is moderate. These are developed countries with an above-average quality of life.
- **Cluster 3**: here are includes countries with the lowest birth rates, very high per capita income, and exceptionally high life expectancy. Also political stability is very strong. These are highly developed and prosperous countries.
- **Cluster 4**: this cluster comprises countries with moderate birth rates, medium per

capita income, and average life expectancy. Political stability is moderate too. These countries are in an intermediate position in terms of socio-economic development.

The figure 9 show a representation of the world map with the highlighted cluster of membership: is a useful visualization to easily understand which cluster a country belongs to and to gain an overview of the division of the level of development in different regions of the world. This type of map can also provide a **geographical and spatial perspective** on the socio-economic characteristics of countries and facilitate the understanding of global development patterns.

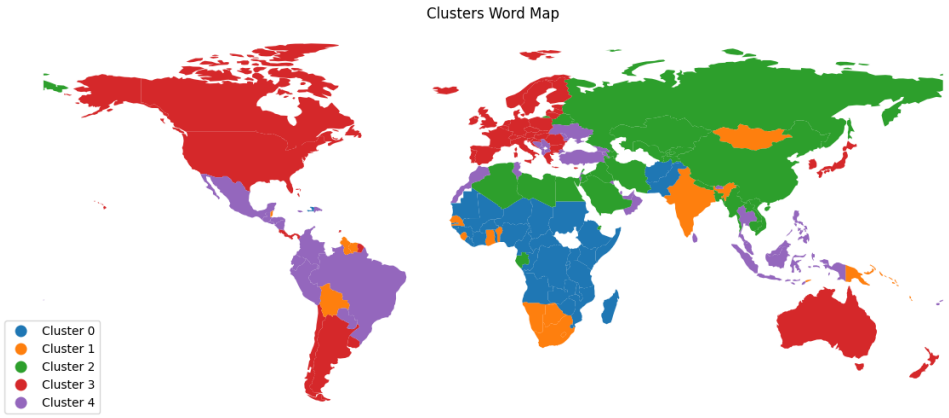

Figure 9: World map

5.4 Relation between Life Index and Clusters

Since the **Life Index** is a synthesis of all the columns we used for computing the clusters, we can use it to perform a comparison between the regions.

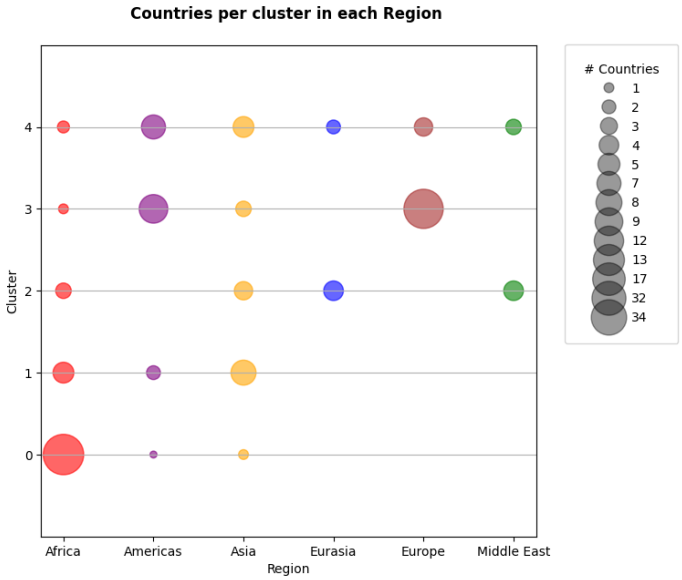

Figure 10: Countries per cluster in each region.

In the figure 10, there is a scatterplot with the cluster of membership on the Y-axis and the region on the X-axis. For each cluster and for each region, the number of countries belonging to them is highlighted through *'bubbles'*, which will be larger, the greater the number of countries in them.

- It can be immediately noticed that **Europe** has only countries from clusters 3 and 4, which identify highly developed or developing regions. Therefore, it appears to be the continent that offers greater security.
- Conversely, the opposite can be said for **Africa**, which provides countries in each cluster, but the majority of them fall into cluster 0, indicating those countries with socio-economic challenges.
- For **Eurasia** and the **Middle East**, the situation is very similar to Europe. There are no underdeveloped countries in both regions, as they are exclusively divided into clusters 4 and 2, indicating either an above-average or intermediate lifestyle.
- For the **Americas**, a significant number of countries are in a condition of development or high development, with a moderate presence of countries facing challenges, notably Haiti.
- For **Asia**, countries are evenly distributed across each cluster, thus presenting the greatest variety of living conditions. This ranges from the high living conditions of Australia (symbolically included in this region) to the development conditions of various archipelagos, to the precarious conditions of Afghanistan and Pakistan.

6 Life Index predictions with ARIMA

6.1 The ARIMA Model

The **ARIMA model, or Autoregressive Integrated Moving Average model**, is a statistical model that is used to forecast future values of a time series. The model is based on the assumption that the past values of the time series can be used to predict future values.

The ARIMA model is comprised of three components:

- *Autoregression (AR)*: it models the relationship between the current value of the time series and its past values.
- *Integration (I)*: it accounts for the non-stationarity of the time series.
- *Moving average (MA)*: it models the random errors in the time series.

From a *mathematical viewpoint*, the model can be expressed by the following equation:

$$y_t = \phi(L)y_{t-1} + \theta(L)\epsilon_t \tag{1}$$

where:

- $y_t$is the value of the time series at time t;
- $\varphi(L)$ is the autoregressive polynomial;
- $\theta(L)$ is the moving average polynomial;
- ϵ_t is the white noise term.

The **autoregressive polynomial** models the relationship between the current value of the time series and its past values. The **moving average polynomial** models the random errors in the time series.

The **order of the ARIMA model** is denoted by the three numbers (p, d, q). The p parameter represents the number of autoregressive terms, the d parameter represents the degree of integration, and the q parameter represents the number of moving average terms. For example, an $ARIMA(1, 1, 1)$ model would have one autoregressive term, one integrated term, and one moving average term.

6.2 Application in this context

In the context of this paper, the ARIMA model was implemented, with the appropriate modifications, in two different situations in order to collect useful insight on the following scenarios:

1. **understanding how the value of Life Index will develop in the next years**. Since this indicator gathers the information from all the pertinent sources analyzed, it represents a meaningful metric in the understanding of the status in which a Country lies.
2. **forecast Life Index features for a single country**.

In particular, the dataset used for these analysis was properly preprocessed to achieve different granularity in data.

6.3 Regional Life Index forecast

This section focuses on the **prediction of the Life Index metric based on regional data aggregation**. The analysis provides a forecast of the Life Index value from an high level overview, highlighting the peculiarities of each region the data points were grouped into. Consequently, it is achievable from the user to engage with an overview on the global Life Index development, encouraging the comparison of the difference projections between

single geographical areas.

First of all, the figure 11 shows the Life Index trend from 2006 to 2021. As it is possible to observe, **the Life index value is not stable** and some downward tendency are clearly visible, as well as upwards. The contrast is more evident if we compare territories that are very diverse from each other, from a political, economic, and social perspective. An example is represented by the *Middle East* and *Europe*, where the first seems to have an overall falling of Life Index value, while the Life Index for latter seems to the on an improving trajectory.

It is worth mentioning that the overall direction is not constant since, on the short period, it displays various peaks or downfalls that can last for a variable period of time before move back to the previous trend. An illustration of this phenomena is represented by the Life Index in *Europe* between 2019 and 2021: a rapid decrease in the overall metric, followed by an increase is observable . On the other hand, the Life index values in the *Americas* seems to not have recovered yet from the downfall began in 2016. Further investigations could be performed to determine whether this is behavior is due to the election of Donald Trump as President of the United States Of America [10].

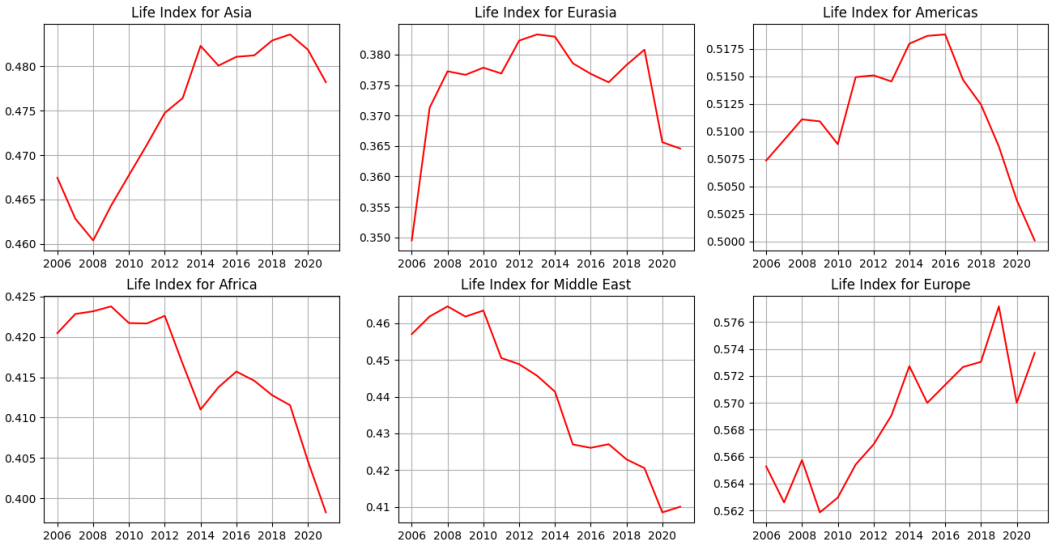

Figure 11: Life Index trend for each region from 2006 to 2021.

Since the goal was to produce Life index predictions for each region an automatic procedure was developed. Therefore, each region had its peculiar model, resulting in the creation of **six different ARIMA models**, tailor made on the specific time series.

6.3.1 ARIMA model's steps

Below are illustrated the general steps that led to the results (showed in section 6.3.2):

Step 1. *Stationarity and Differencing (if necessary)*: time series model needs **stationary data**: stationary data refers to the time series data in which mean, variance, and autocorrelation, do not vary across time. If data is not stationary, we have to transform it by **differencing**, which means removing trends and seasonality.
However, ARIMA (AutoRegressive *Integrated* Moving Average) models are designed to work with non-stationary data, and the **"Integrated"** component involves differencing the data to achieve stationarity.
Stationarity is an important assumption in many time series modeling techniques, including ARIMA, as it ensures that patterns observed in the past can be expected to continue into the future.
In the ARIMA model the "integrated" component is denoted by the parameter d.
The **Dickey-Fuller Test** is a statistical test used to determine whether a given time series is stationary or not and it is often used in econometrics and time series analysis. The test involves formulating **null** and **alternative hypotheses**:

- *Null Hypothesis (H_0)*: the null hypothesis assumes that the time series has a unit root, meaning it is non-stationary. In other words, it has a trend or some form of dependence that causes it to vary over time.
- *Alternative Hypothesis (H_1)*: the alternative hypothesis assumes that the time series is stationary, indicating the absence of a unit root.

In simpler terms:

- **If p-value < significance level (commonly 0.05)**: reject the null hypothesis, indicating stationarity.
- **If p-value ≥ significance level**: fail to reject the null hypothesis, indicating non-stationarity.

Step 2. *Model Identification*: plotting the **Autocorrelation Function (ACF)** and **Partial Autocorrelation Function (PACF)** on the differenced data is useful to identify potential values for the autoregressive (AR) and moving average (MA) components of the ARIMA model.
The plots then needs to be intrepreted. Below a general guideline to interpret the ACF and PACF plots:

- *ACF (Autocorrelation Function) Plot*: the ACF plot shows the correlation between a time series and its lagged values. It helps you **identify the potential MA order (q)** of your ARIMA model. If the ACF plot decays exponentially to zero, it suggests a non-stationary time series that requires differencing (d). On the other hand, If the ACF plot has a sharp drop after a certain lag and then tails off, it suggests an MA process of order q (order of moving average).

– *PACF (Partial Autocorrelation Function) Plot*: the PACF plot shows the correlation between a time series and its lagged values while controlling for the effects of intermediate lags. It helps you **identify the potential AR order (p)** of your ARIMA model. If the PACF plot has a sharp drop after a certain lag and then tails off, it suggests an AR process of order p (order of autoregressive).

The plots shown in figure 12 provide the plots described above, helping in a more deeply understanding of the model and its parameters, with respect to the specific characteristic of the single time series.

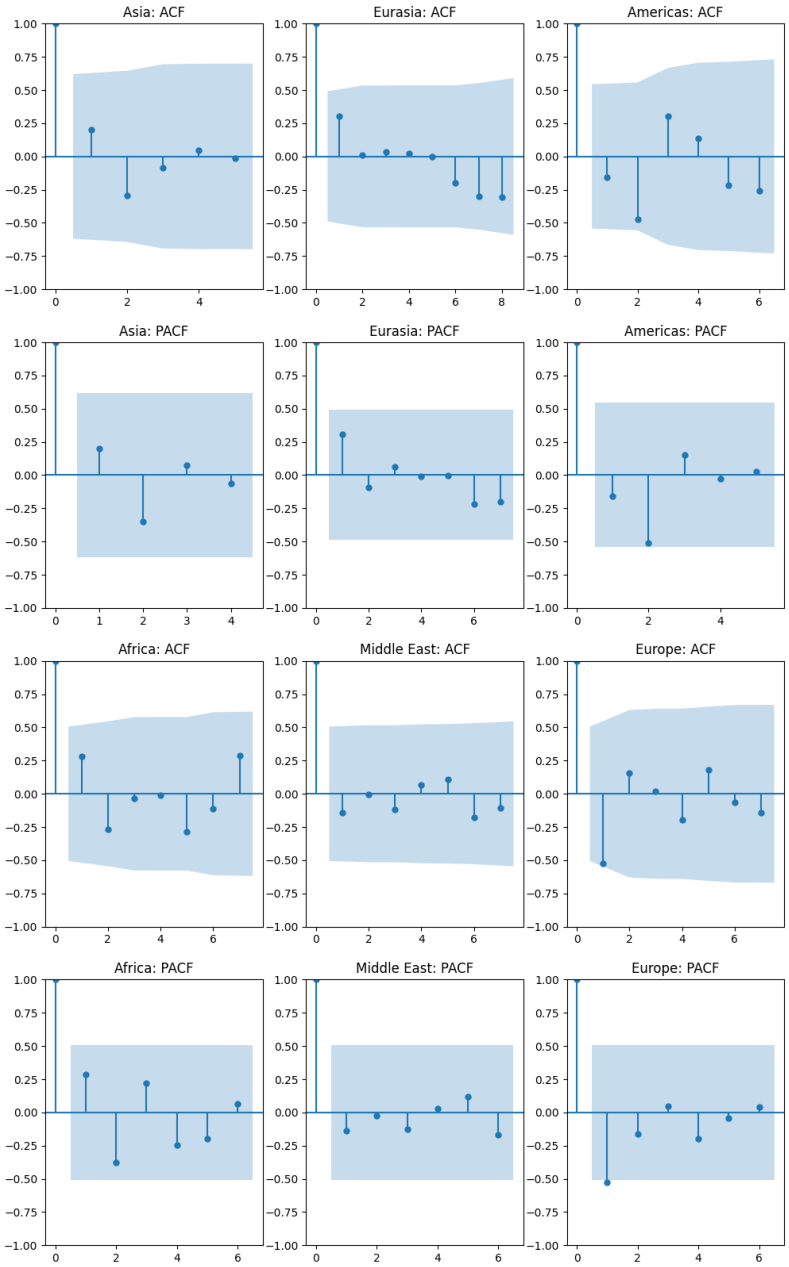

Figure 12: ACF and PACF plots for each region.

Step 3. Model Estimation and model fitting: it is possible to use the information gained from the ACF and PACF plots to select the values for the ARIMA parameters (p, d, q). However, since we would like to create and automatic procedure, we implemented the **automatic method** `auto_arima` from the `pmdarima` library.
In order to find the best model, `auto_arima` optimizes for a given `information_criterion`, in this case the **Akaike Information Criterion** *aic*, and returns the ARIMA which *minimizes* the value. Nevertheless, the value found in Step 1 for the parameter d was manually inserted into the model, since it usually returned better results.

The best parameters for each each region are listed below:

Region	p	d	q
Middle East	0	1	0
Europe	1	1	0
Eurasia	1	0	0
Asia	1	3	0
Americas	0	2	1
Africa	0	1	1

Table 3: ARIMA paramaters per region.

The dataset is not split into training set and test set, since the number of data points is not large enough to guarantee efficient results. Hence, all the data is is given as input to the model.

Step 4. **Forecasting**: future values of the time series are predicted.

6.3.2 Results

Regarding the predictions, the **number future steps was equal to 5**, meaning that the Life Index value **from 2022 to 2026** was forecast.

The figure 13 exhibits the original data and the predicted ones, incorporating also the **confidence interval levels**. In this case, a 95% was built, signifying that there is a 95% probability that the true value of the next observation will fall within the interval. The wider the confidence interval, the less certain the prediction is. A wide confidence interval means that there is a higher probability that the true value of the next observation will fall outside the interval, while the narrower the confidence interval, the more certain the prediction is. A narrow confidence interval means that there is a lower probability that the true value of the next observation will fall outside the interval.
In general, a 95% confidence interval is considered to be a good level of confidence.

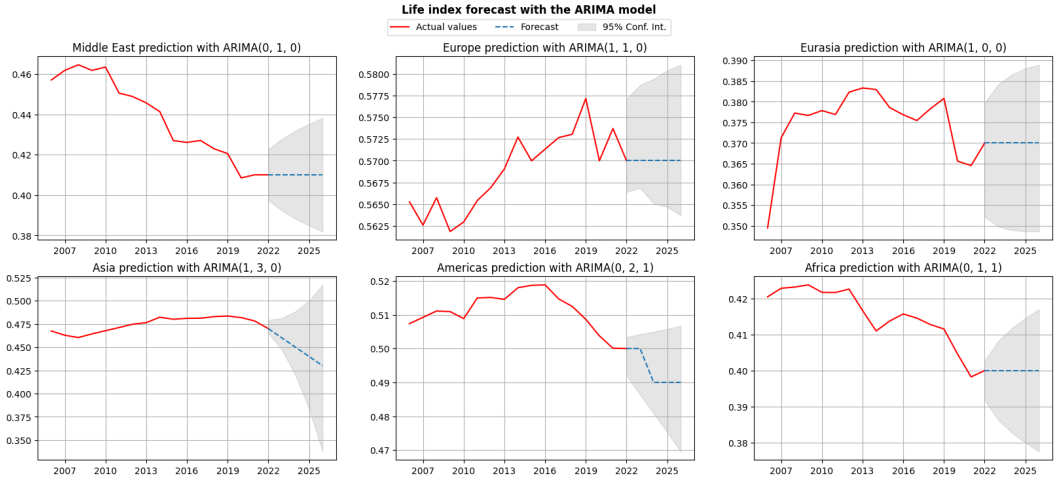

Figure 13: Life Index forecast for Countries grouped by regions.

In our context, the ARIMA model(s) is usually able to detect the trend from the data, returning a valid overview on how the Life Index will develop in the next years, even if, for some regions, we observe a static situations, that is the model predicts a stationary value, and also enlarges the confidence interval.
Regardless, the results are generally consistent with previous findings and warrant further investigation.

6.4 Country's Life Index feature forecast

The study aims to emphasize the previous forecast on the Life Index metric based on regional data aggregation by **implementing a forecast on the features that form the index**. Therefore, in this case, only the data from a single country is considered.
Each field is forecast to the year 2026, with the intention of focusing on the situation of the single Country, knowing that it could resemble the one observed on the specific region the Country belongs to or diverge from the latest observation.
In particular, the ARIMA model predicts the *denormalized values* of the features that are subsequently normalized and summed in order to recreate the Life Index. In this way, it is possible to gain a better understanding on the single country situation and the resulting index.

In the matter of the model used tor this task, also in this case an ARIMA model was implemented. The model undergone all the steps described in Section 6.3.1.

From this point forward, **Italy** is the country chosen for the analysis but the remarks made in the next paragraphs are valid for all the countries under study. The data employed in the analysis are shown the table 4.

As can be observed, the values through the years tend to not be constant, since upwards and downwards trends are observable in the various time series. For example, while *Life Expectancy* (`Life exp.`) appears to have increased in recent years, *Birth rate* is reducing,

Edition	Total	GDP	Pol. Stab.	Birth rate	Life exp.	Life index
2006	0.92	32451.493	0.531	9.6	81.283	0.578
2007	0.92	34145.546	0.447	9.7	81.434	0.577
2008	0.92	35523.271	0.548	9.8	81.485	0.584
2009	0.9	34603.256	0.347	9.6	81.637	0.570
2010	0.89	35156.668	0.474	9.5	82.037	0.576
2011	0.89	36598.015	0.501	9.2	82.188	0.578
2012	0.89	36486.297	0.508	9.0	82.239	0.578
2013	0.88	36314.697	0.495	8.5	82.69	0.575
2014	0.9	36194.874	0.458	8.3	83.09	0.578
2015	0.89	36899.385	0.376	8.0	82.544	0.569
2016	0.89	39926.955	0.369	7.8	83.244	0.575
2017	0.89	41581.121	0.307	7.6	82.946	0.572
2018	0.89	43036.244	0.342	7.3	83.346	0.576
2019	0.89	45799.772	0.404	7.0	83.498	0.582
2020	0.89	43144.406	0.428	6.8	82.195	0.572
2021	0.9	46705.018	0.578	6.8	82.795	0.588

Table 4: Italy's data.

confirming Italy's falling birth rate [11]. It is noteworthy that, despite the observed fluctuations in the data, Life Index is on a fairly steady increasing trend, suggesting that the country is the overall performing nicely over the aspects measured for the study.

6.4.1 Results

In total, **5 ARIMA models** were created and implemented, one for each feature selected: *GDP*, *Total*, *Political Stability*, *Birth rate* and *Life expectancy*. The table 5 reports the best parameters that were chosen for each time series.

Feature	p	d	q
GDP	0	1	0
Total	1	0	0
Political Stability	0	0	0
Birth rate	1	3	0
Life expectancy	1	1	0

Table 5: ARIMA parameters for each feature in Life Index (Italy).

Regarding the forecast value, as mentioned before, the data for the year 2026 was predicted, that is 5 years after the latest data point available. The table 6 show the results.

Edition	Total	GDP	Pol. Stab.	Birth rate	Life exp.	Life Index
2026	0.900	46705.017	0.444	8.987	82.615	**0.413006**

Table 6: Italy's forecast at 2026.

The value predicted by the models seems coherent with the original data and provide a **useful representation** of the guess on the future situation of the country.
The table also reports the computed Life Index value computed on the normalized values from the features, indicating a slightly decrease in the general status of the country. This could be due to the fact that some variables values predictions are lower than the original ones, penalizing the final score. However, despite the changes occurred, the country situation

is foresee to remain stable. In fact, the results do not suggest that the country, in the next years, will alter the cluster in which it was previously allocated, ensuring a high well-being for its inhabitants.

7 Conclusions

In this project, we developed *Life Index*, a scalar metric that ranges from 0 to 1, to measure and delineate the quality of life within a given country. We used historical data from the past 15 years to predict the evolution of all indicators, providing a look into the future direction of nations around the world. We also used clustering methodologies to uncover patterns and similarities between nations based on these key indicators. This information can be used to develop policies and programs that can help to improve the well-being of people around the world.

It is important to note that the Life Index is just one measure of well-being. Other factors, such as access to clean water and sanitation, personal safety, and social support, can also play a role in determining the quality of life and might deserve further investigations. For this reasons, it might make sense to integrate these new sources of information into the analysis, in order to gain a deeper understanding of the situation in a given country.

Considering a longer time horizon, one can aim for significantly better results by increasing the effectiveness of predictions. This could help in various domains by providing greater security. This project has the potential to be used in a variety of practical applications, such as helping people choose where to live, work, or invest. For example, a government could use the Life Index to track the progress of its country's development, a business could use the Life Index to identify countries with attractive investment opportunities, a an individual could use the Life Index to choose a country to live in or visit.

References

[1] Freedom in the Word

[2] Freedom 2023 Dataset

[3] World Bank Organization

[4] World Bank Data Platform

[5] World Bank Data Population

[6] World Bank Data GDP per capita

[7] World Bank Data Political Stability and Absence of Violence/Terrorism

[8] World Bank Data Birth Rate

[9] World Bank Data Life Expectancy

[10] 2016 United States presidential election

[11] Italy's falling birth rate