### **Introduction**

Government education expenditure is a key indicator of a country’s commitment to education and human capital development. A well-trained workforce is a long-term investment that can significantly boost a nation’s economy, reduce unemployment rates, and increase national productivity as gross domestic product (GDP).

In this project, we aim to use historical data on countries’ education expenditures to build statistical and machine-learning models that can forecast future education expenditures, explain factors related to education expenditure, and prescribe strategies for long-term economic prosperity.

## Data Collection

We utilized three datasets for this project:

1. **Government Expenditure as Percentage of GDP**

- Source: [World Bank]

(<https://data.worldbank.org/indicator/SE.XPD.TOTL.GB.ZS?end=2021&start=1999>)

2. **Countries’ GDP Per Capita**

- Source: [World Bank]

(<https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2021&start=1999>)

3. **Countries’ Unemployment Rate**

- Source: [World Bank]

(<https://data.worldbank.org/indicator/SL.UEM.TOTL.NE.ZS?end=2021&start=1999>)

The primary dataset for this project is the Government Expenditure dataset, while the countries’ GDP per Capita and Unemployment Rate datasets serve as supplementary datasets.

## Methods and Tools

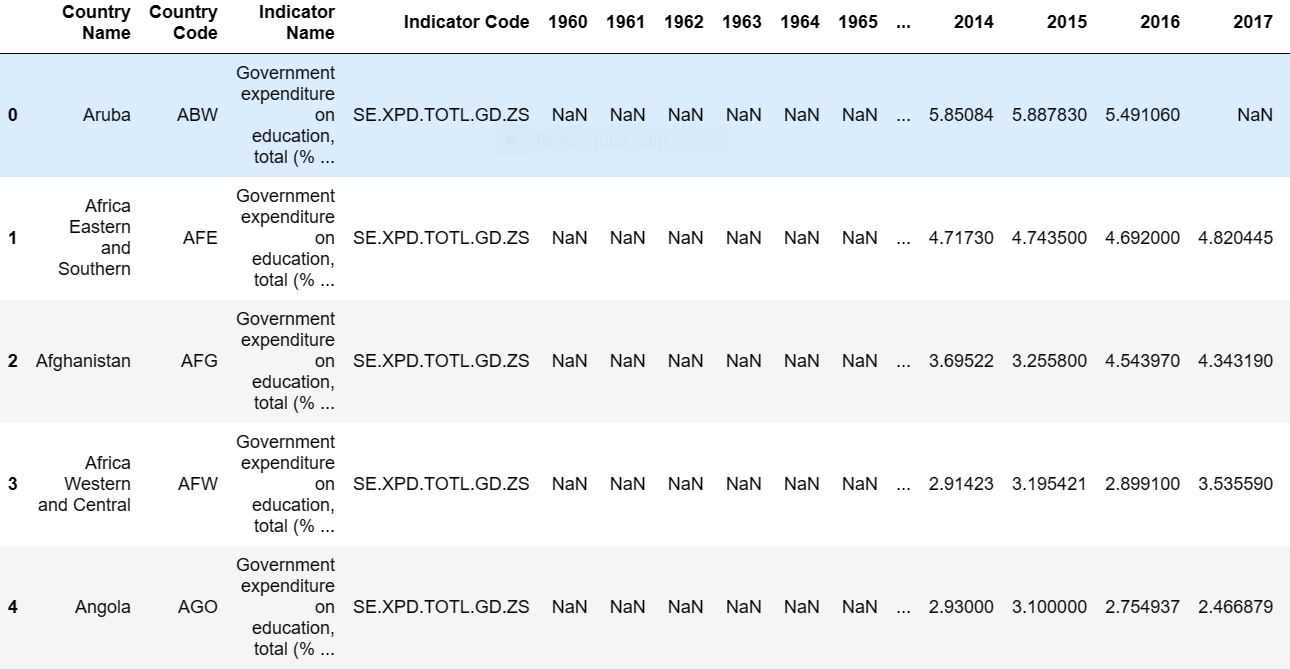
We used the Python programming language for the entirety of our project. The following packages and versions were utilized.

* Tensorflow : 2.10.0
* Pandas : 2.0.3
* Numpy : 1.25.0
* Scikit learn : 1.2.2
* Matplotlib : 3.7.2
* Seaborn : 0.12.2

## Data Cleaning

We loaded the raw Government expenditure dataset into a Pandas Dataframe. The raw dataset has 68 columns and 266 rows.

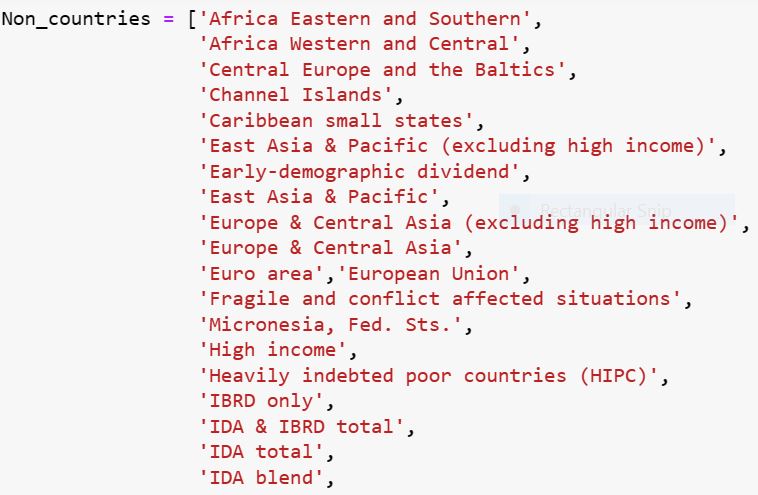
Preview of the raw dataset:



**Removal of Irrelevant Data**

We identified the key features and data points needed for the analysis and dropped all irrelevant columns.

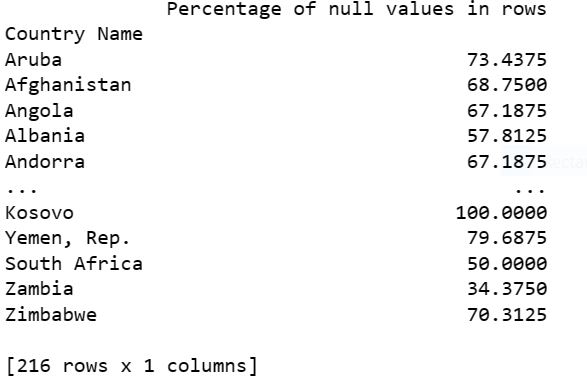
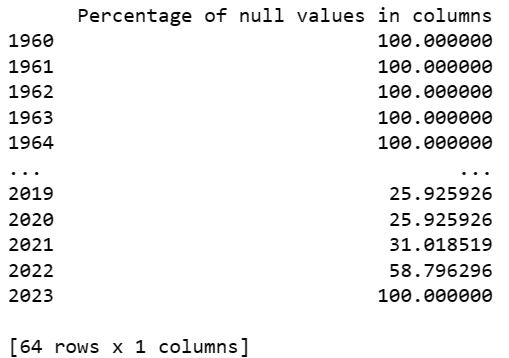
Preview of a list of irrelevant rows to be dropped:



### **Handling Missing Values**

Exacly 64 percent of the entries in the dataset were missing or null. Due to the need for strictly sequential data to execute an effective time series analysis, we dropped all missing entries. After dropping missing values, we had a dataset with 23 columns representing 23 years of historical data and 19 rows representing 19 different countries.

Preview of percentage of null values in the dataset:



### 

### 

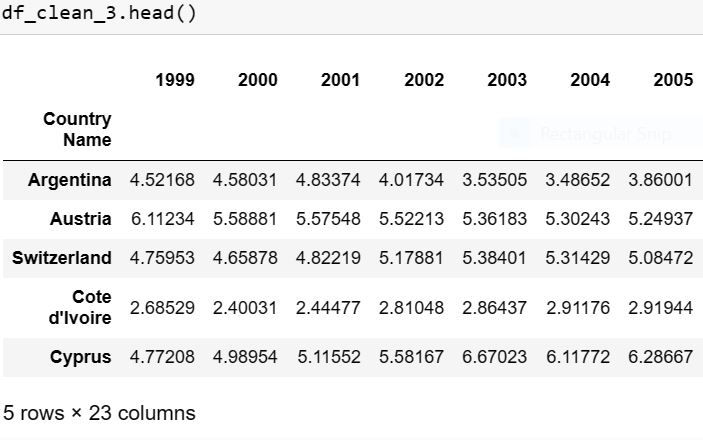
### 

### Data Transformation

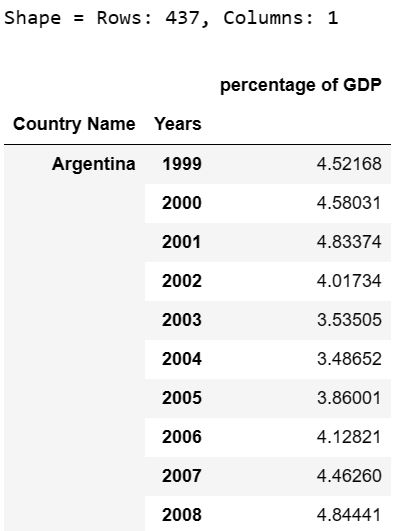
To enable statistical and machine-learning analysis, we transformed the dataset by:

1. Converting the string columns to Pandas period objects.
2. Stacking (using the stack method) the DataFrame to convert it from the wide form to the long form. This conversion created a DataFrame with two indices, one for the country and the other for the period (years), and a single column for government expenditure.

Preview of cleaned dataset before transformation:



Preview of cleaned dataset after transformation:



*The above data-cleaning steps were also applied to the secondary datasets.*

## 

## 

## 

## 

## 

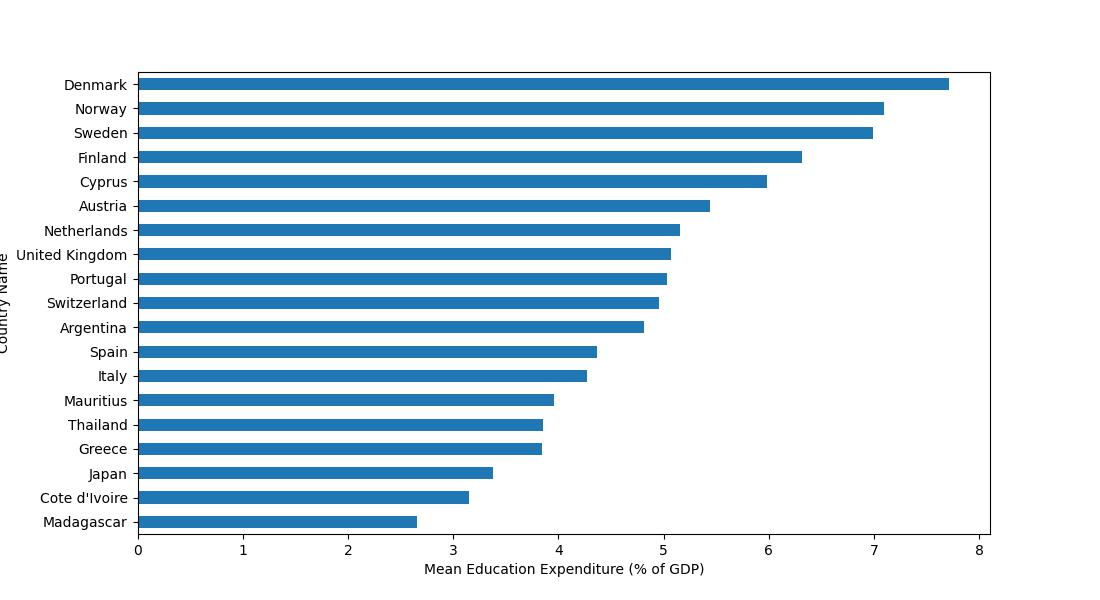
## 

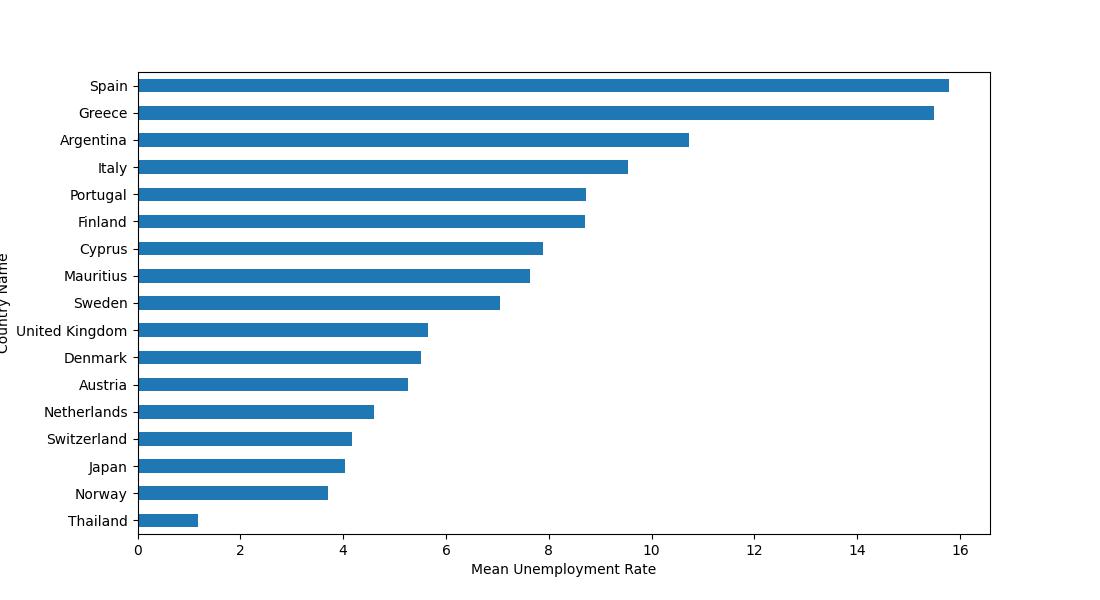
## 

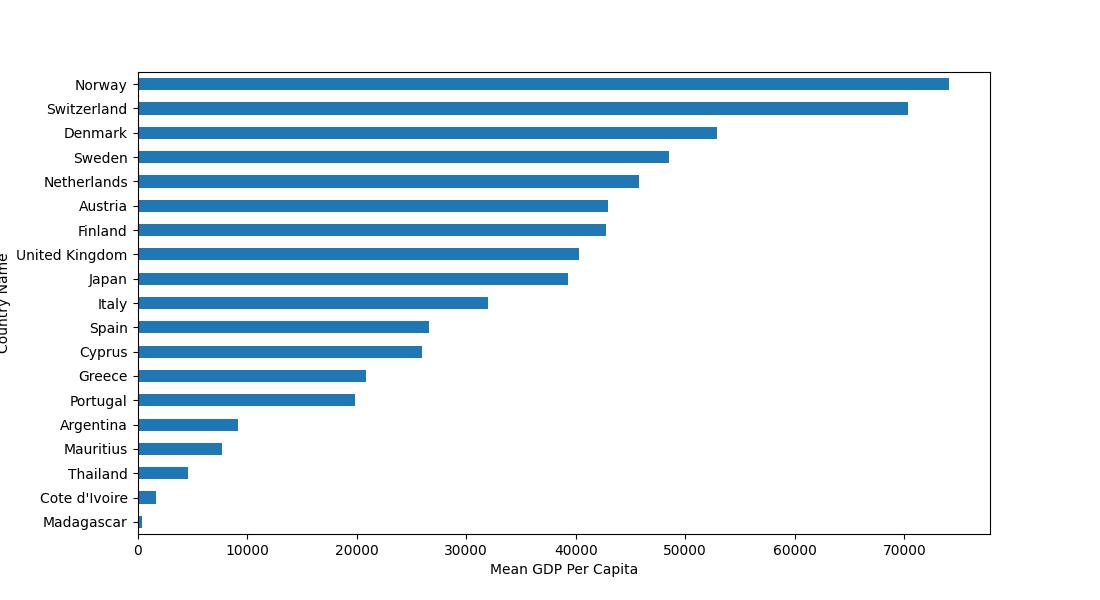
## **Exploratory Data Analysis**

### **Descriptive Statistics**

We grouped and aggregated the dataset by country name and obtained the mean and median of education expenditure, unemployment rate, and GDP per capita over 23 years for each country. We produced bar charts to rank countries according to their education expenditure, unemployment rate, and GDP per capita.

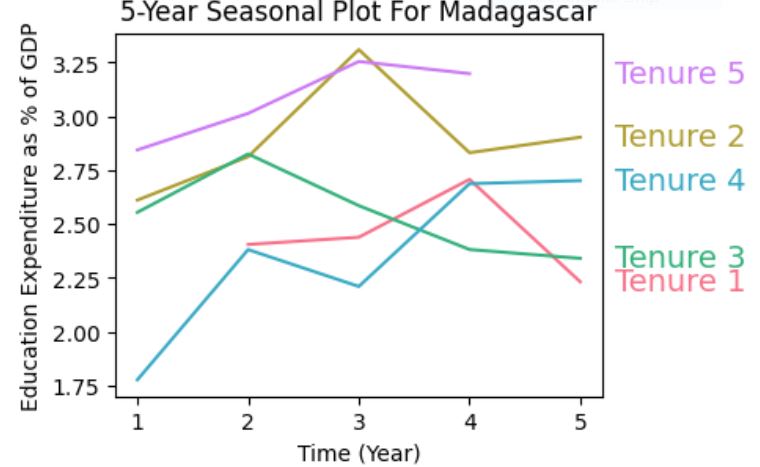
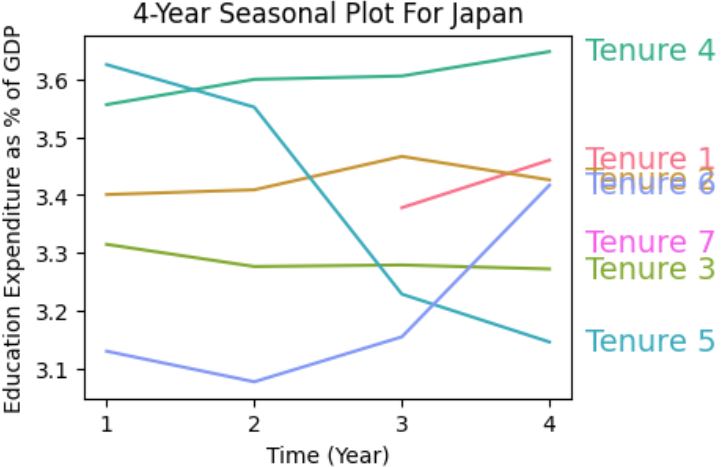
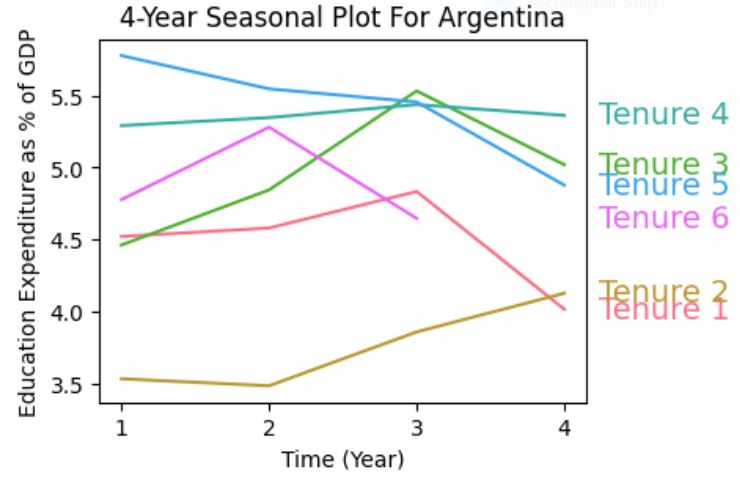
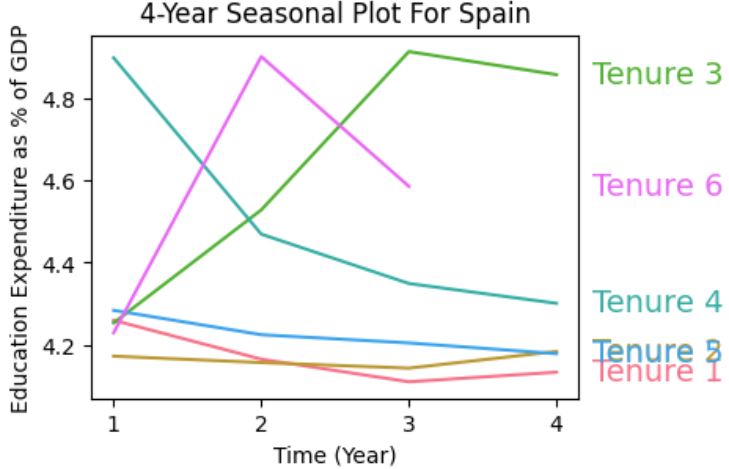






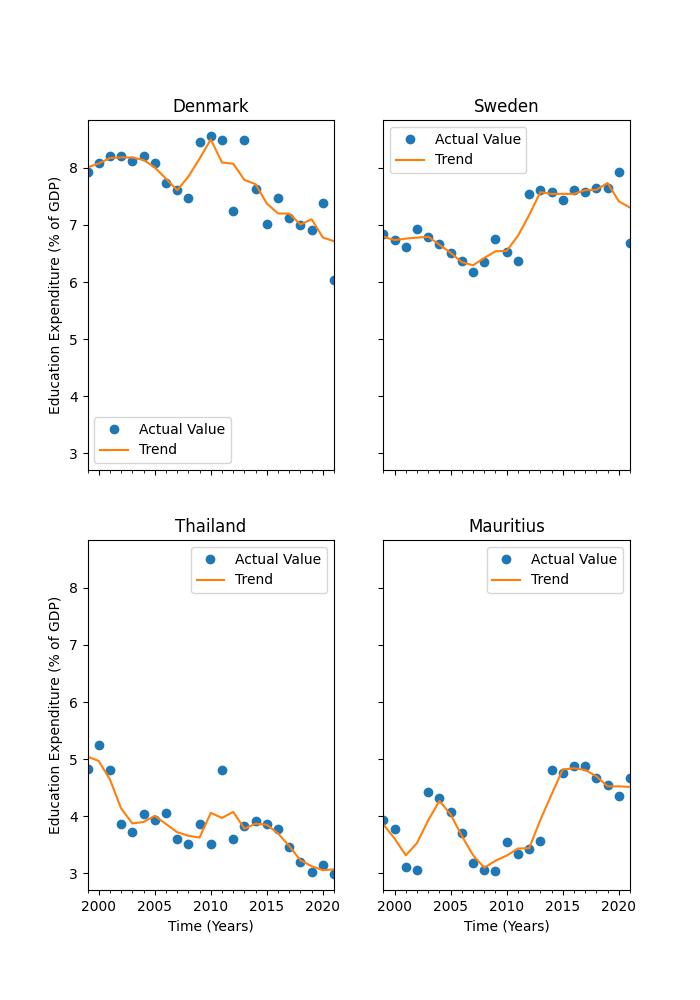
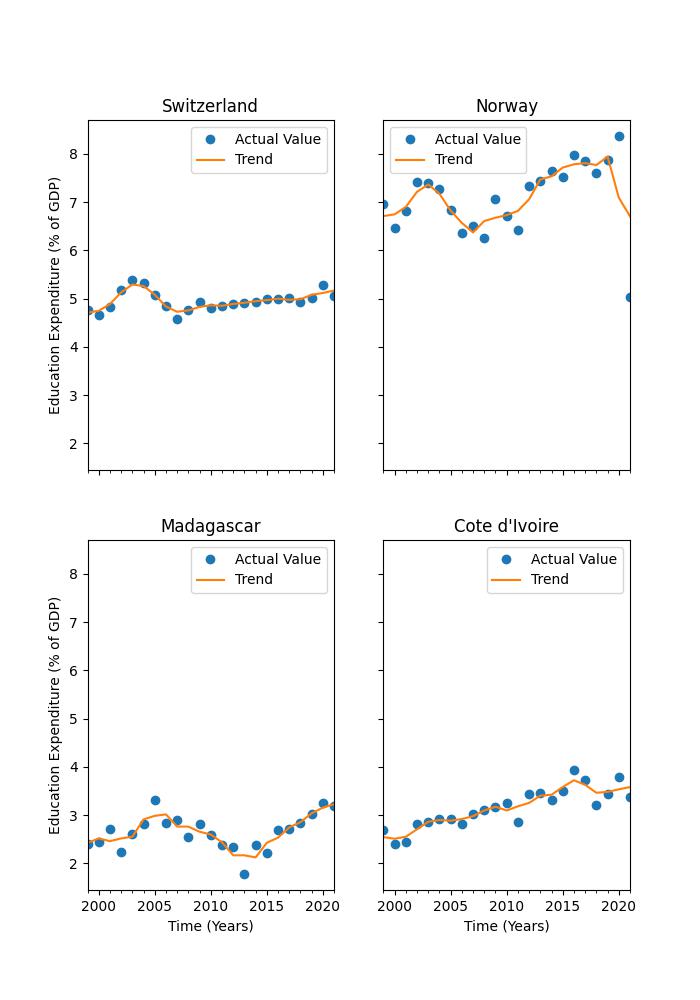
### **Seasonality Analysis**

We created seasonal plots following the electoral/government cycles of each country. For instance, a seasonal plot for Argentina considers Argentina’s 4-year election cycle and actual election years. There were no persistent seasonal patterns in education funding in all the countries. We concluded that educational expenditure did not have a seasonal pattern.

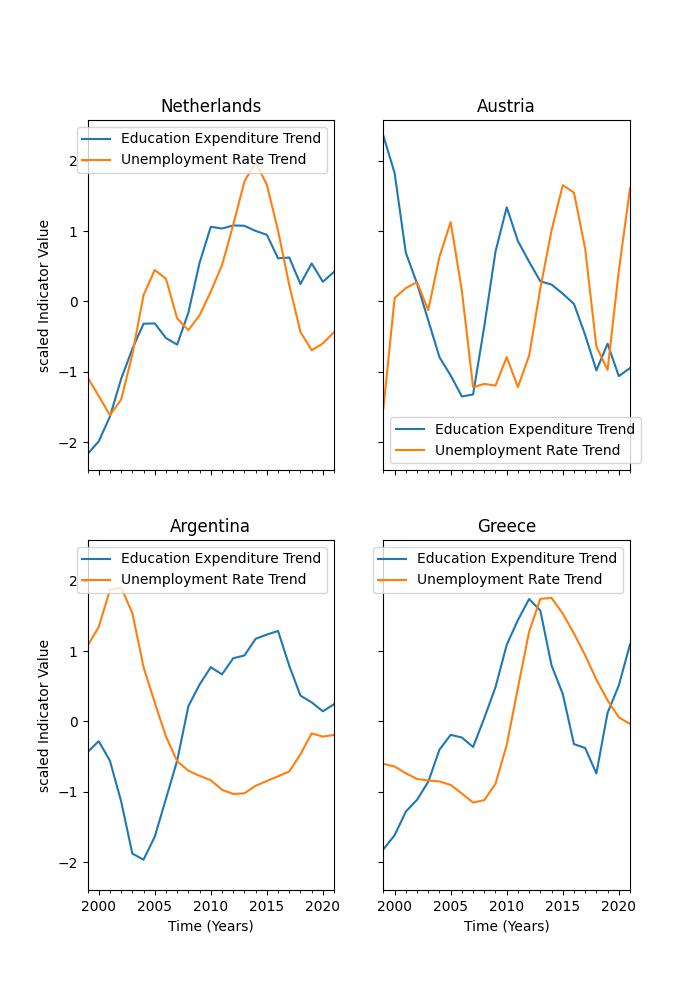
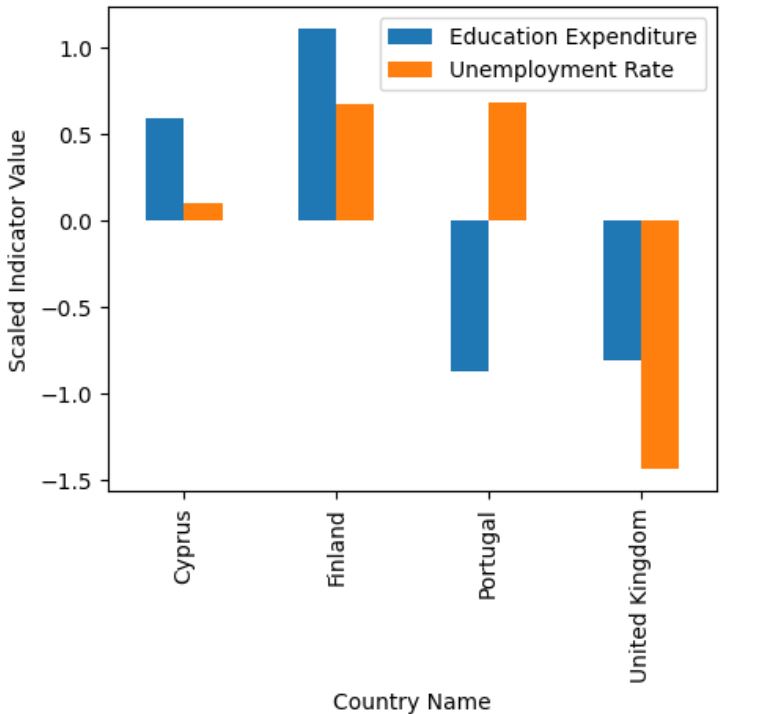


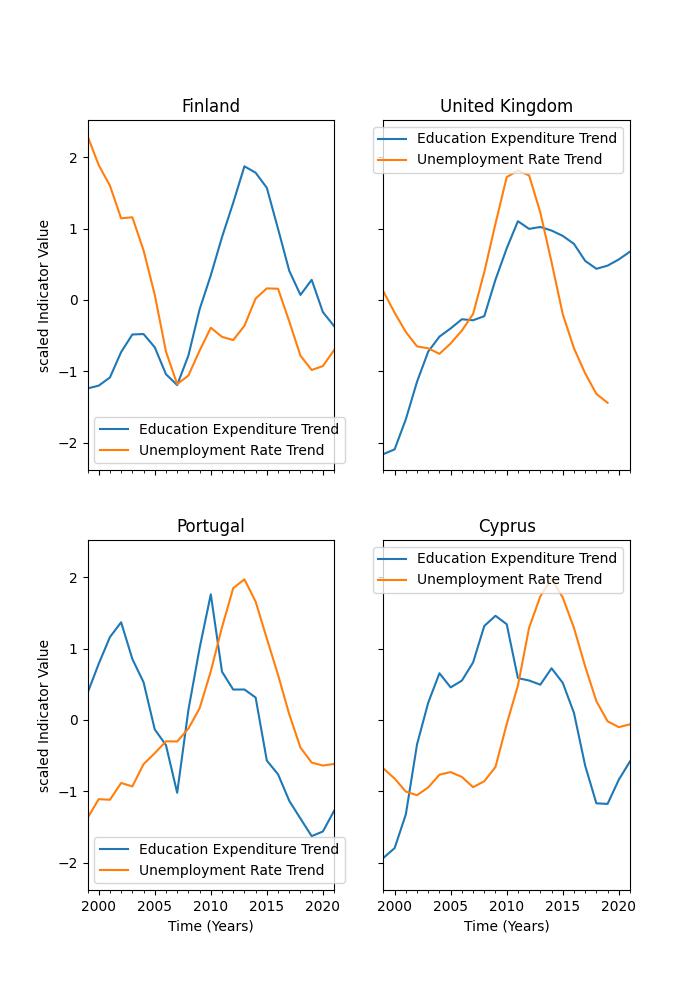
### **Trend Analysis**

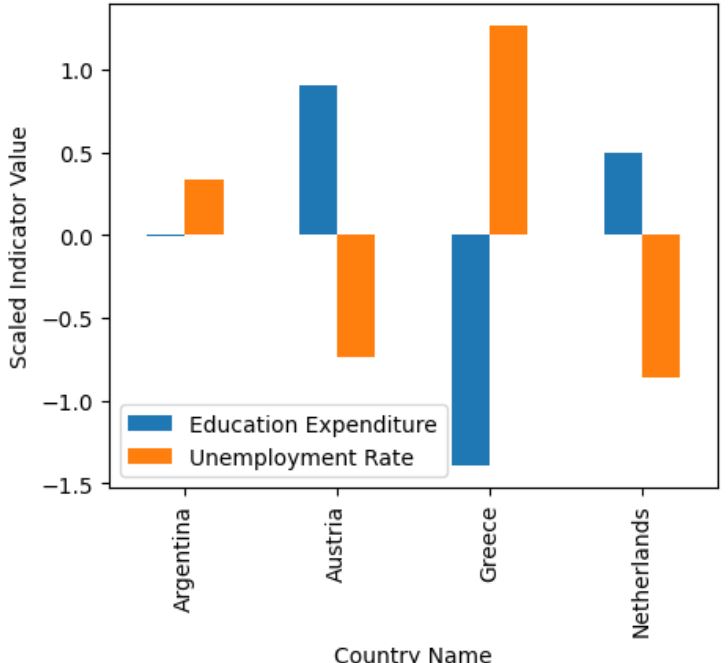
We generated trend line plots to explore long-term changes in the mean of key economic indices for each country.



To study the complex relationship between education expenditure and other key economic indicators, we created comparative trend graphs (line graphs and bar graphs) comparing unemployment rates with educational expenditures on the same scale. We observed a predominantly inverse relationship between government spending on education and unemployment rates. In other words, as governments increase their expenditure on education, their unemployment rates reduce, and vice versa, irrespective of the country’s GDP per capita.







There were no relevant correlations between GDP per capita and education expenditure. Some wealthy countries had a declining trend in education expenditure (e.g., Denmark), while some poorer countries were increasing their education budgets (e.g., Ivory Coast). However, this relationship is entirely arbitrary.

### **External Events Affecting Government Expenditure**

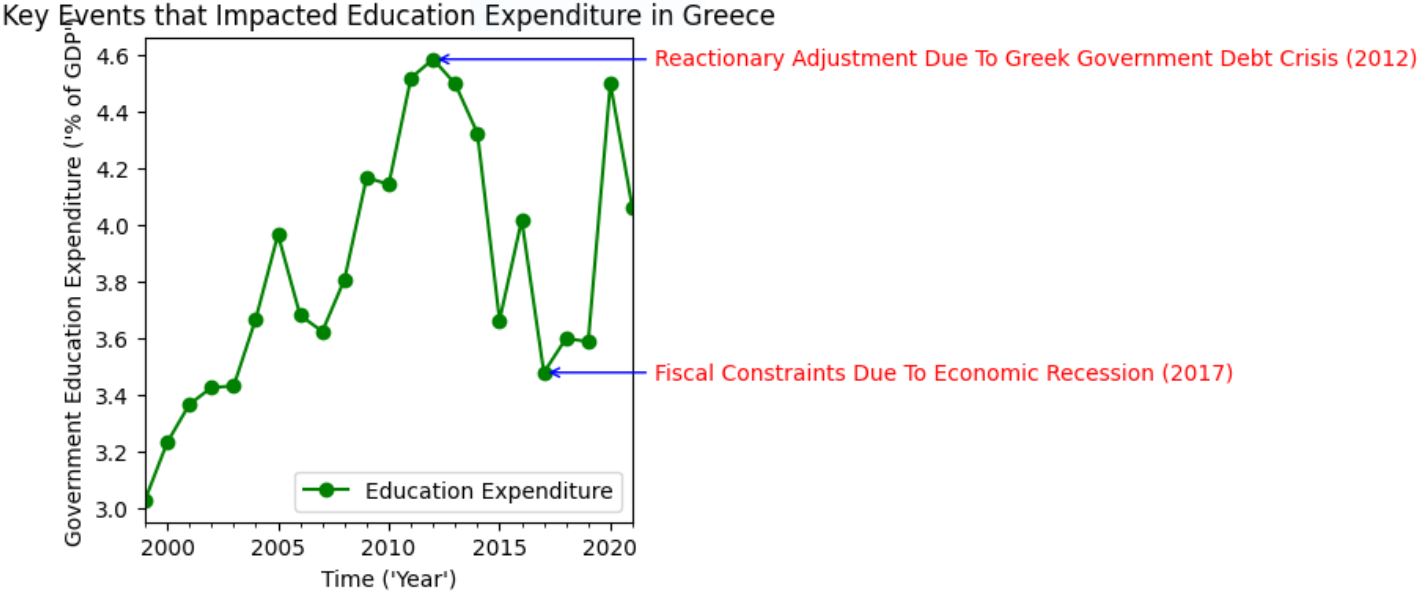
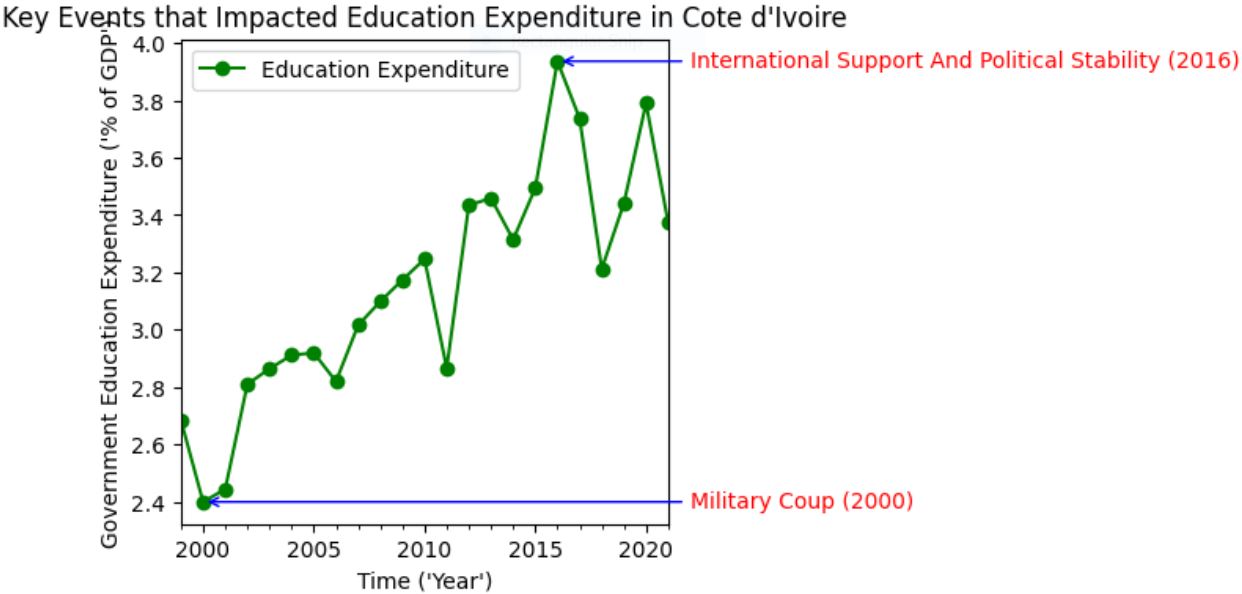
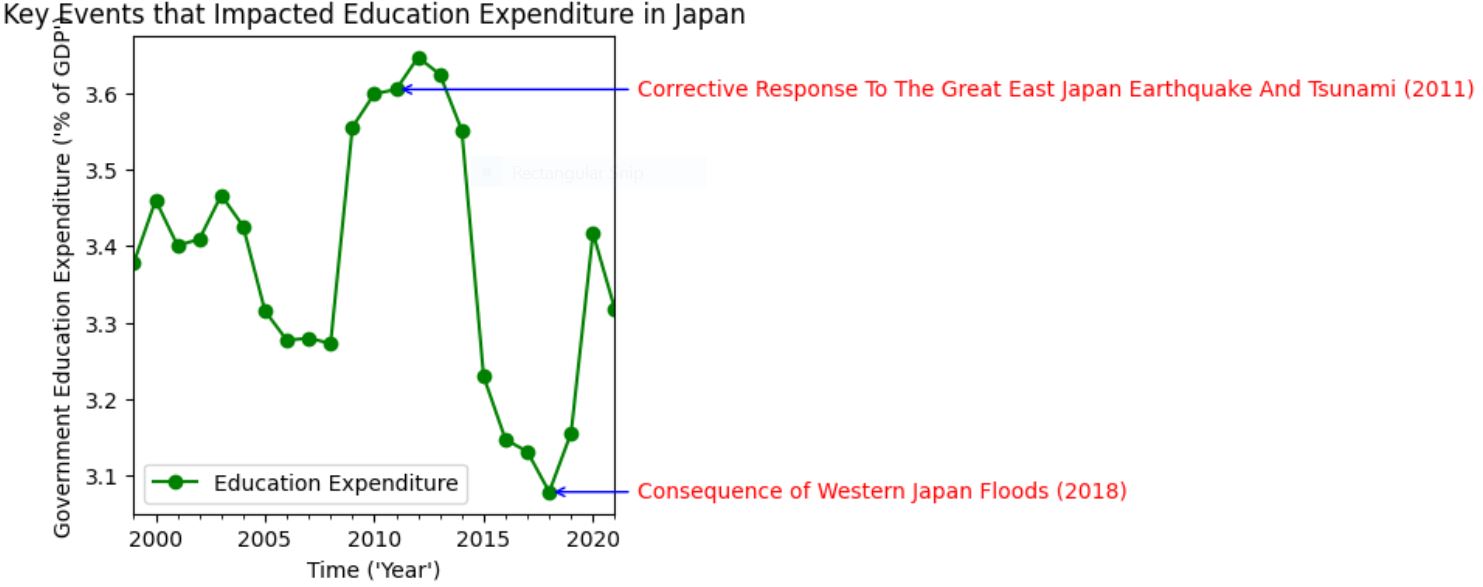
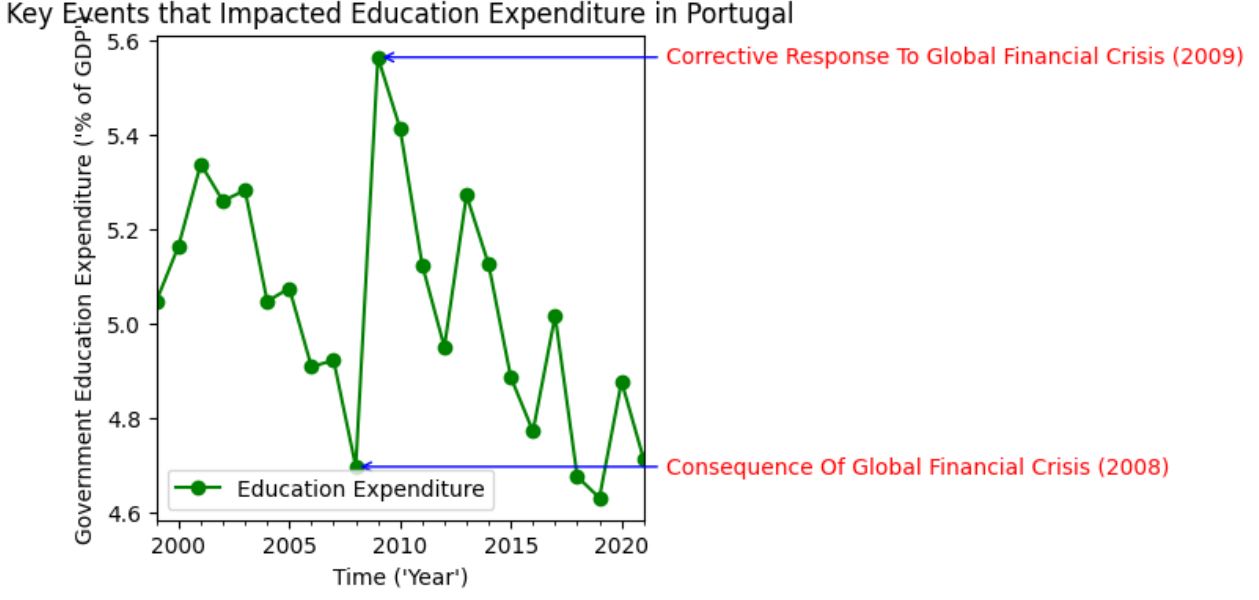
Several factors that cannot be measured or observed graphically influenced government education expenditure in the 19 countries observed between 1999 and 2021. These factors include:

1. **Economic Recessions**: Some countries may respond to economic recessions by increasing expenditure to stimulate economic recovery, while others may reallocate their education budget to other critical areas.

2. **Natural Disasters**: Countries may shift focus from education to other areas such as healthcare, social welfare, and environmental protection after a major environmental crisis like a tsunami, earthquake, or pandemic.

3. **International** **Aid**: When countries receive international financial support, they are often advised to invest in human capital development by increasing their budget on education.

4. Political interest, war, and conflicts.



### **Data Splitting**

The cleaned and preprocessed time series dataset was split into training and testing sets to enable model evaluation. The dataset consisted of 23 years of historical data for 19 different countries. The first 20 years of data for all 19 countries were used as training data, while the last 3 years were used as test data. All 19 training datasets were shuffled after splitting to eliminate bias.

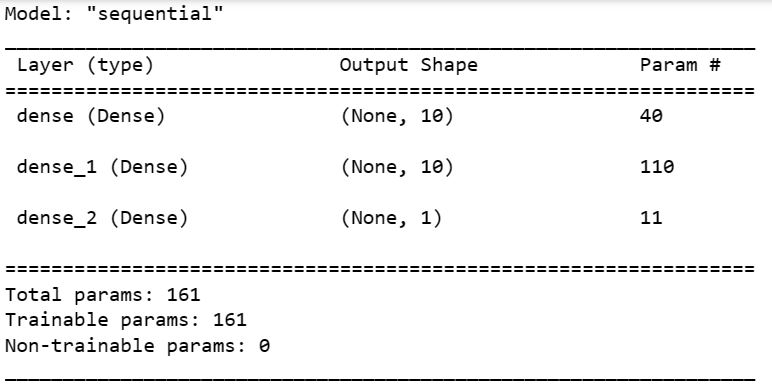
### **Feature Engineering for Machine Learning**

The lack of seasonality observed during EDA necessitated the use of lagged variables as features for the time series machine learning. We used the Window method of TensorFlow’s Data API. Lagged variables are simply the values of the target variable at preceding time steps. For instance, if a target variable is measured in 2024, three-year-lagged features will be values recorded in 2021, 2022, and 2023. Lagged variables capture the relationship between historical data and present data and leverage that relationship in predicting future data. After extensive experimentation, we utilized 3-step lagged variables as our features.

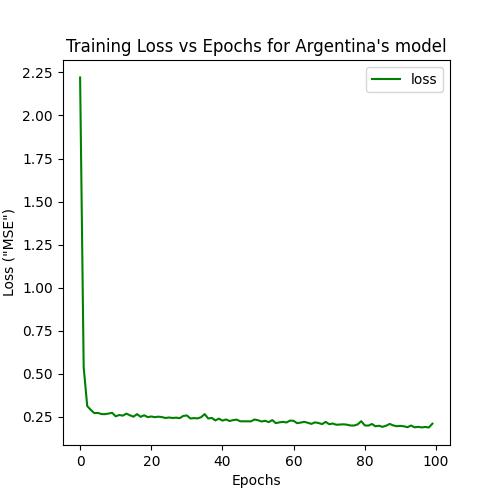
**Model Selection and Training**

Due to the small size of the dataset, extensive model selection was not feasible. We selected a three-layer deep neural network model with 10 neurons in the first and second layers and a single neuron in the output layer, also with a total of only 161 parameters. ReLU activation functions were used in the first and second layers, while the final output layer used a linear function. Each model was trained for 100 epochs using the Adam optimizer at a learning rate of 0.001 and Mean Square Error (MSE) as the loss function. The different models trained for each country were stored in a Python dictionary for efficient retrieval when needed.

A preview of the model summary:



A preview of the training curve of a sample model:



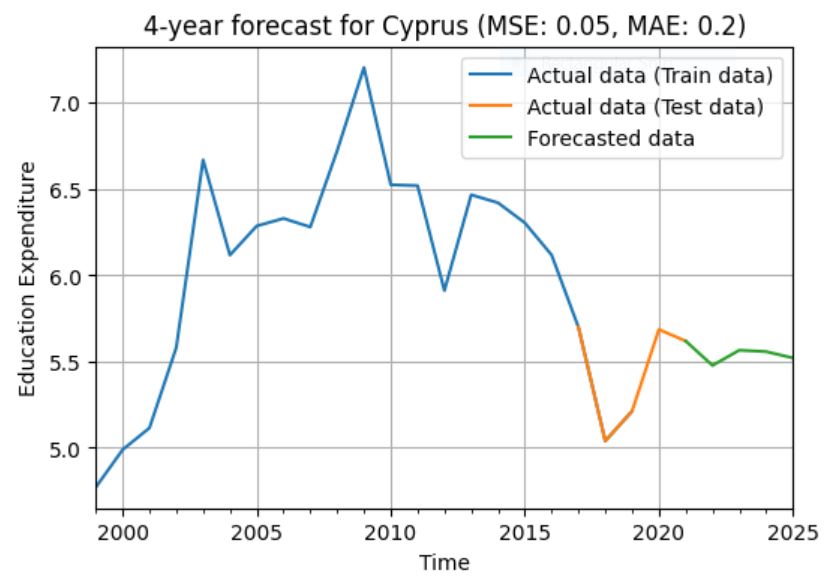
**Model Evaluation**

Each model was evaluated on the test data using Mean Absolute Error (MAE) and Mean Square Error (MSE). All test evaluations for all 19 models produced values between 0.01 - 0.3.

**Predictions**

A custom Python function (`Plot\_Forecast`) was written to enable easy forecasting and graphing of forecasts.

Preview of the time series model output:



**Recommendations**

Governments should proactively address unemployment and economic downturns by investing strategically in education and human capital development at all levels. Reactive measures taken only in response to rising unemployment rates tend to be inefficient and less effective.