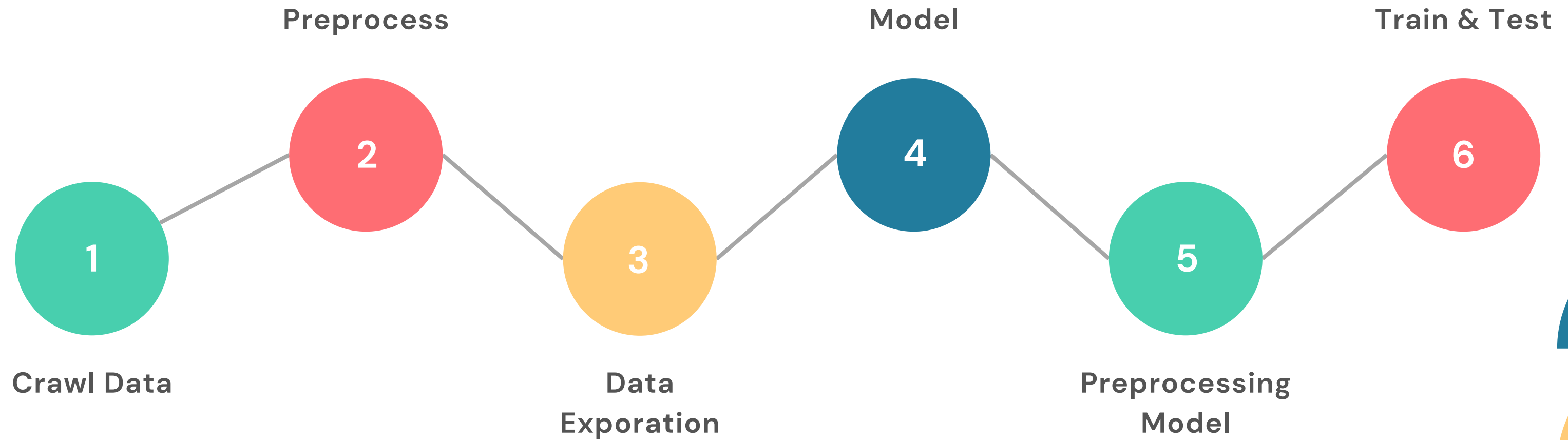# PROJECT

Giảng viên hướng dẫn:
Lê Ngọc Thành
Nguyễn Ngọc Thảo
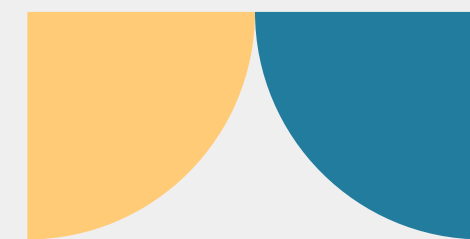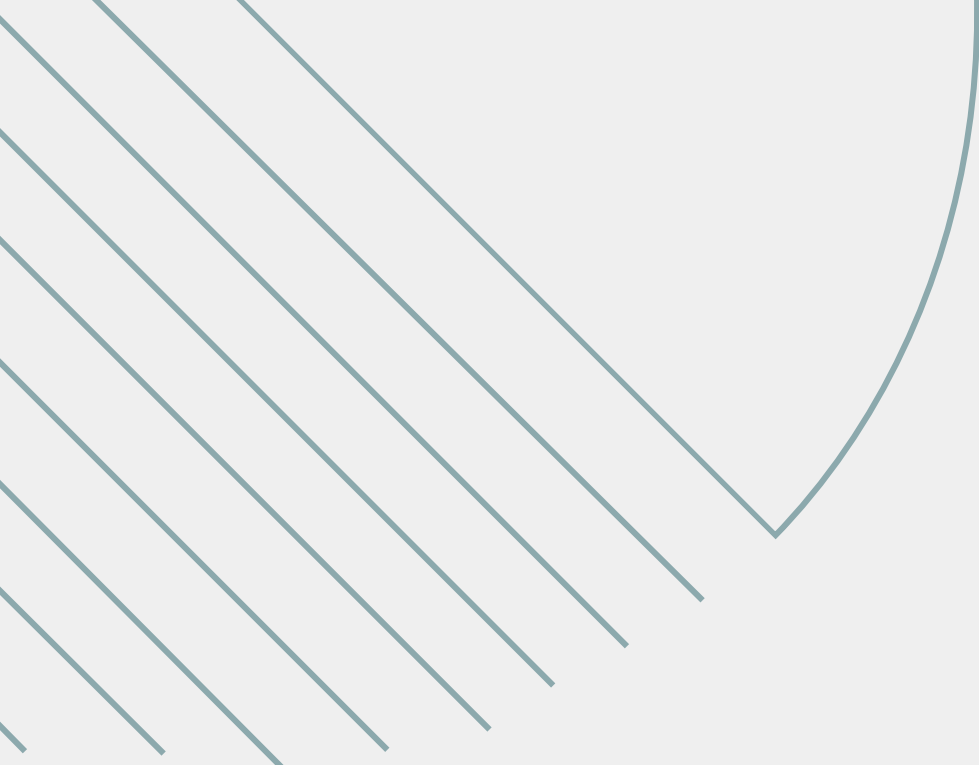Nguyễn Bảo Long
Phạm Trọng Nghĩa

THÀNH VIÊN

- Lương Trường Thịnh – 19127559
- Trần Trọng Tín – 20127683
- Lê Nguyễn Minh Quang – 20127295
- Nguyễn Đỗ Nguyên Phương – 21127399

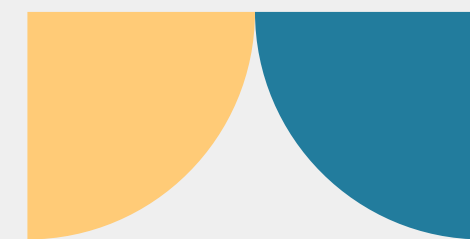# MỤC LỤC

Crawl Data 1

Preprocess 2

Data Exporation 3

Model 4

Preprocessing Model 5

Train & Test 6

CRAWL DATA

CRAWL DATA

- BASE_URL = 'http://api.worldbank.org/v2/'
- Region: Asia
- Year: 2015:2022

PREPROCESS
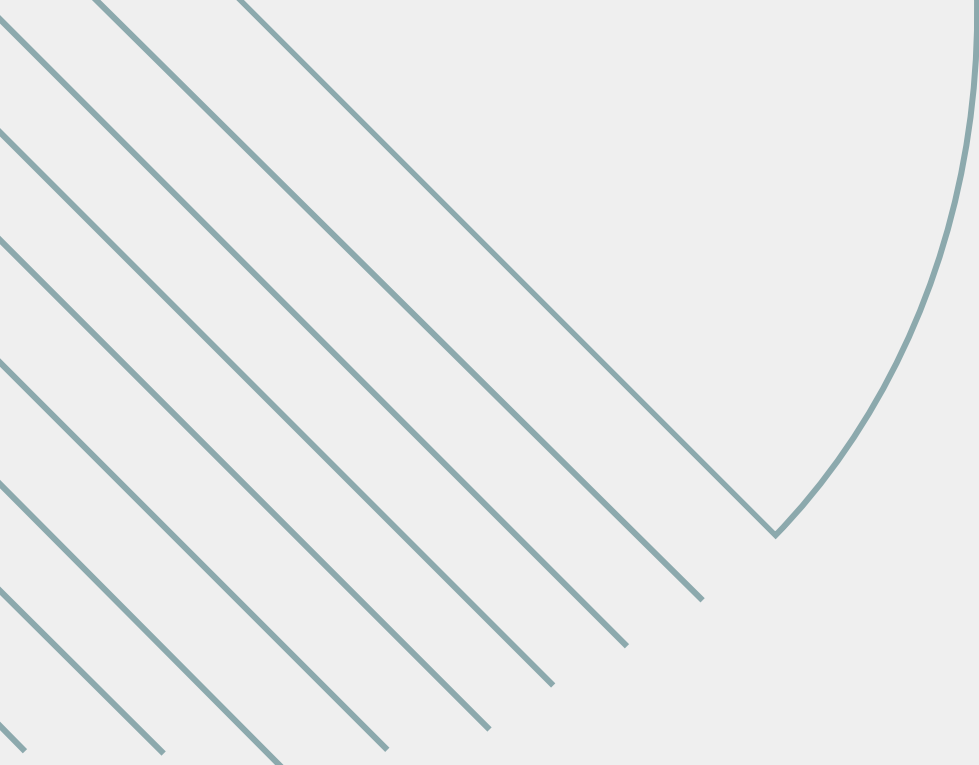
PREPROCESS

Preprocessing data involves manipulating data to make it suitable for modeling:

- Standardization of data: Ensuring stability and uniformity in the data's type.
- Handling missing data: Removing missing data such as null or None values
- Noise removal: Eliminating unnecessary or irrelevant data or noise.

DATA EXPLORATION

BASIC LEVEL

Answer these questions :

- How many rows and columns are there in your data?
- What is the meaning of each row/column?
- What is the datatype of each column?
- Is this suitable datatype for the column?
- What is the distribution of the data in each column?

BASIC LEVEL

- How many rows and columns are there in your data?

BASIC LEVEL

- What is the meaning of each row/column?

For row

Shows the data corresponding to each column of the data table

For column

Total Population: The total number of people in the country.

Female Population: The number of females in the country.

Male Population: The number of males in the country.

Birth Rate: The number of live births per thousand of the population per year.

Death Rate: The number of deaths per thousand of the population per year.

Compulsory Education Dur.: The number of years of compulsory education.

Employment in Industry(%): Percentage of the workforce employed in the industry sector.

Employment in Agriculture(%): Percentage of the workforce employed in the agriculture sector.

Female Employment in Agriculture(%): Percentage of female workforce employed in the agriculture sector.

Female Employment in Industry(%): Percentage of female workforce employed in the industry sector.

Unemployment(%): The percentage of the workforce that is unemployed.

GDP in USD: Gross Domestic Product measured in US dollars.

National Income per Capita: The income earned by each individual in the country.

Net income from Abroad: The net income received from foreign sources.

Agriculture value added(in USD): The value added in agriculture sector in US dollars.

Electric Power Consumption(kWH per capita): Electricity consumed per capita.

Renewable Energy Consumption (%): Percentage of energy consumed from renewable sources.

Fossil Fuel Consumption (%): Percentage of energy consumed from fossil fuels.

Male life expectancy: Average life expectancy for males.

Female life expectancy: Average life expectancy for females.

School enrollment, primary: Enrollment rate in primary education.

School enrollment, tertiary: Enrollment rate in tertiary education.

Primary completion rate: Rate of primary school completion.

Literacy rate: The percentage of the population that can read and write.

Year: The year the data was recorded.

Country: The name of the country.

BASIC LEVEL

- What is the datatype of each column?

BASIC LEVEL

- Is this suitable datatype for the column?

Is this suitable datatype for the column?

These data types seem suitable for the corresponding columns, as they match the nature of the data they hold.

Numerical data is stored as floats or integers, and categorical data like country names is stored as objects.

This should allow you to perform various calculations, analyses, and visualizations effectively on this dataset in Python using pandas and other related libraries.

BASIC LEVEL

What is the distribution of the data in each column?

```
       Total Population  Female Population  Male Population  Birth Rate  \
count       1.316000e+03       1.316000e+03     1.316000e+03  1316.000000
mean        9.020217e+07       4.433594e+07     4.586623e+07    20.669686
std         2.549317e+08       1.240024e+08     1.309465e+08     7.693830
min         2.582080e+05       1.249510e+05     1.332570e+05     5.100000
25%         4.382818e+06       2.121473e+06     2.311564e+06    15.272250
50%         1.822701e+07       9.013542e+06     9.061772e+06    19.841000
75%         6.135766e+07       3.051728e+07     3.080141e+07    24.321000
max         1.417173e+09       6.915285e+08     7.311805e+08    52.073000


       Death Rate  Compulsory Education Dur.  Employment in Industry(%)  \
count  1316.000000               1316.000000                1316.000000
mean      6.338262                  8.931611                  21.354252
std       2.669952                  1.797727                   7.752873
min       0.795000                  5.000000                   3.519346
25%       4.876500                  9.000000                  16.338200
50%       6.208000                  9.000000                  20.982230
75%       7.502750                  9.000000                  25.454523
max      16.700000                 15.000000                  59.578700


       Employment in Agriculture(%)  Female Employment in Agriculture(%)  \
count                   1316.000000                          1316.000000
mean                      28.955947                            31.696589
std                       21.396788                            26.516102
min                        0.324730                             0.007847
25%                        6.612762                             3.846060
50%                       29.100450                            30.145570
75%                       44.913765                            51.072530
max                       85.412960                            89.413740
```

```
       Female Employment in Industry(%)  ...       GDP in USD  \
count                       1316.000000  ...     1.316000e+03
mean                          12.886698  ...     4.320544e+11
std                            6.665878  ...     1.456643e+12
min                            1.773412  ...     2.904910e+08
25%                            8.027576  ...     1.258728e+10
50%                           11.879490  ...     5.296849e+10
75%                           16.000835  ...     2.346714e+11
max                           39.333190  ...     1.796317e+13


       Agriculture value added(in USD)  Renewable Energy Consumption (%)  \
count                     1.316000e+03                       1316.000000
mean                      2.955903e+10                         19.764871
std                       1.080030e+11                         26.125644
min                       4.972527e+07                          0.000000
25%                       1.006268e+09                          1.197500
50%                       3.638505e+09                          5.285000
75%                       1.865373e+10                         33.080000
max                       1.311311e+12                         94.370000


       Male life expectancy  Female life expectancy  \
count           1316.000000             1316.000000
mean              68.710672               74.103446
std                6.157865                6.028307
min               51.039000               54.087000
25%               64.592000               70.332750
50%               68.479000               74.716000
75%               73.142750               78.256500
max               83.100000               87.710000


       School enrollment, primary  School enrollment, tertiary  \
count                 1316.000000                  1316.000000
mean                   101.367796                    31.898819
std                     10.492194                    19.338474
min                     22.162991                     0.212900
25%                     97.937523                    19.731844
50%                    100.911263                    28.845509
75%                    104.800289                    38.867870
max                    150.354233                   125.763786


       Primary completion rate  Literacy rate          Year
count              1316.000000    1316.000000   1316.000000
mean                 94.780845      97.773745   2008.500000
std                  10.378208       4.807482      8.080818
min                  17.885321      46.990051   1995.000000
25%                  95.512396      98.709351   2001.750000
50%                  96.636715      98.709351   2008.500000
75%                  97.566250      98.709351   2015.250000
max                 134.545609     100.000000   2022.000000
```
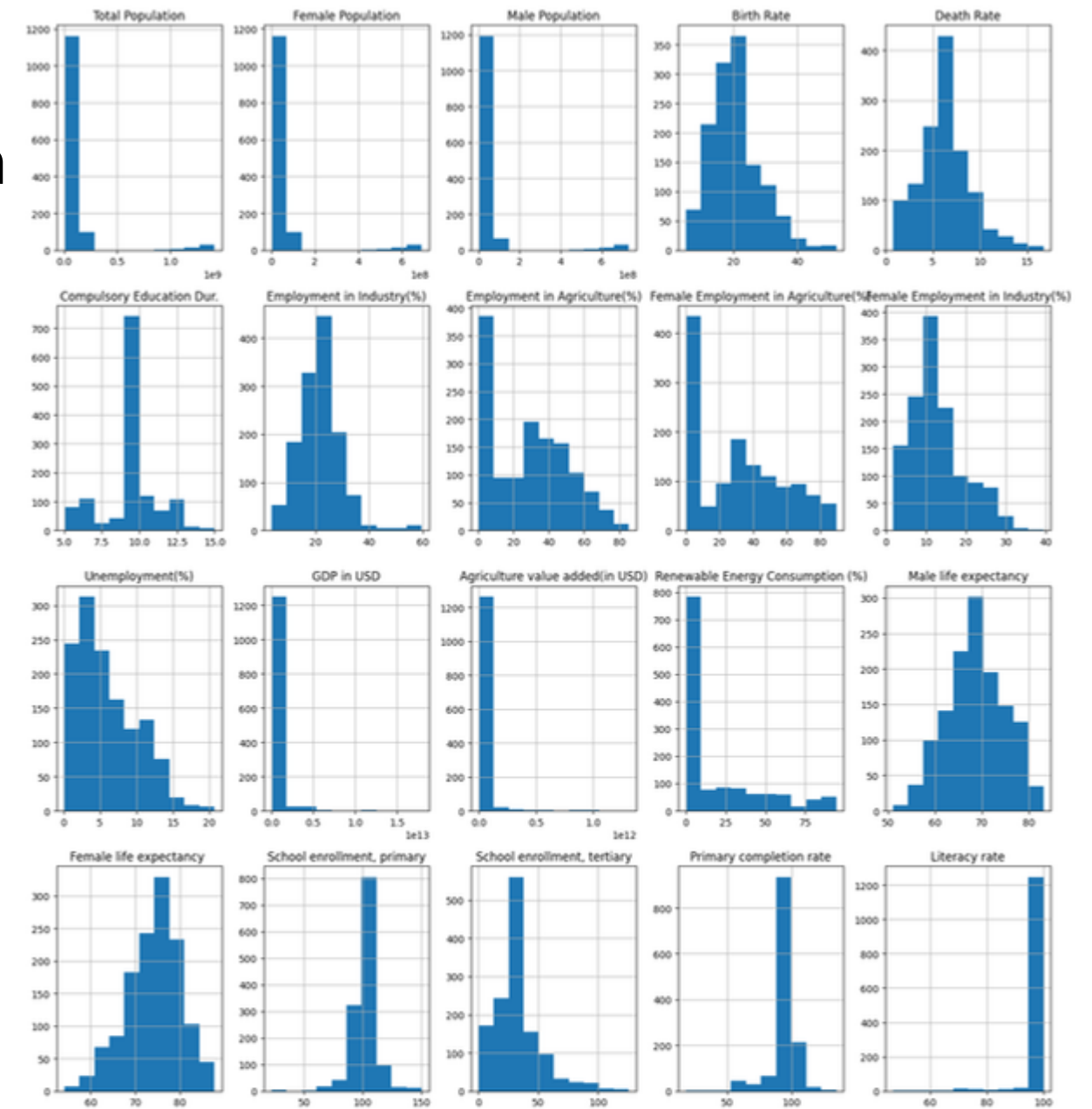
BASIC LEVEL

The chart displays layout the distribution of the data
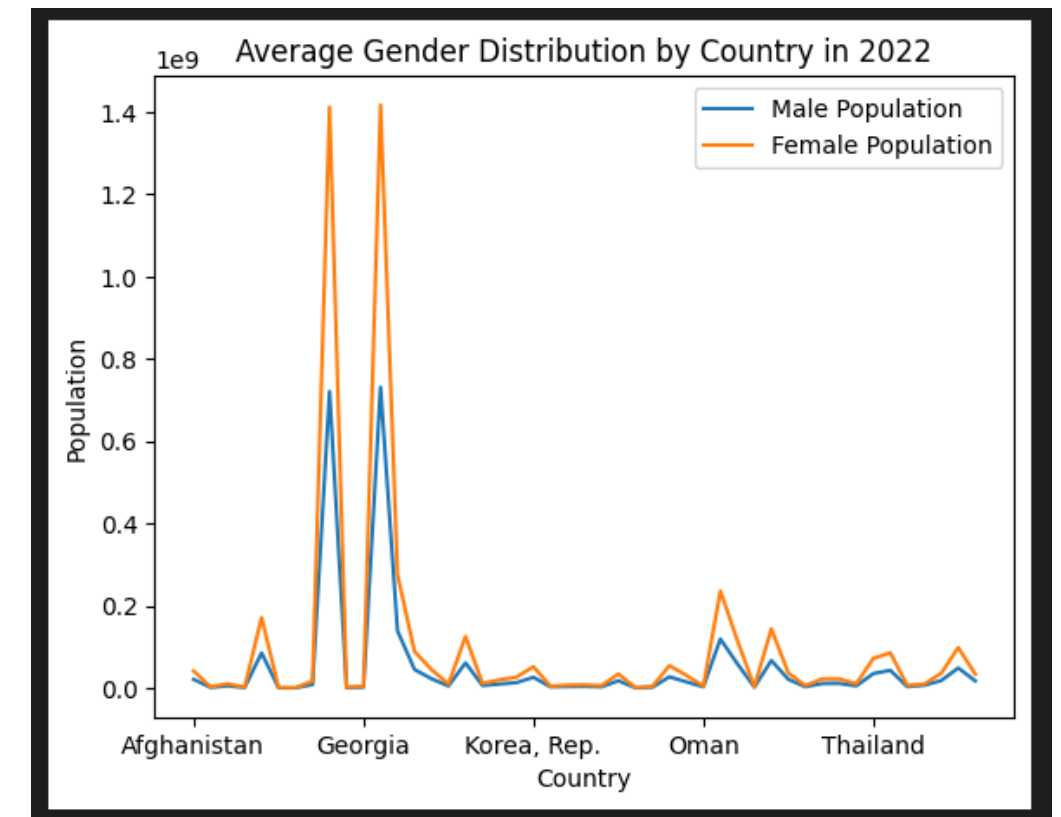
MAKE OWN QUESTION

- Đặt những câu hỏi để tìm hiểu rõ hơn về data

Making your own question 01? (1.5 points)

A possible question is : How does the gender distribution vary by continent?

Answering this question will: show the differences in male and female populations across continents.

How we answer this question: Calculate the average male and female populations for each continent.

MAKE OWN QUESTION

- For more understanding about our data

Making your own question 02? (1.5 points)

A possible question is : Is there a correlation between a country's GDP and its birth rate growth in 2020?

Answering this question will: help us understand the relationship between economic development and population growth.

How we answer this question: Calculate the correlation between GDP and population growth for each country in 2020.

| Country	GDP in USD	Birth Rate
Afghanistan	2.014345e+10	36.1
Armenia	1.264170e+10	12.5
Azerbaijan	4.269300e+10	12.5
Bahrain	3.462181e+10	12.6
Bangladesh	3.739022e+11	18.1
Bhutan	2.325186e+09	12.6
Brunei Darussalam	1.200580e+10	14.2
Cambodia	2.587280e+10	19.8
China	1.468774e+13	8.5
Cyprus	2.500827e+10	10.6

MAKE OWN QUESTION

- For more understanding about our data

Making your own question 03? (1.5 points)

A possible question is : How does the gender distribution life expectancy by continent in 2020?

Answering this question will: show the differences in male and life expectancy across continents.

How we answer this question: Calculate the average male and female life expectancy for each continent.

Average Gender Distribution by Country in 2022
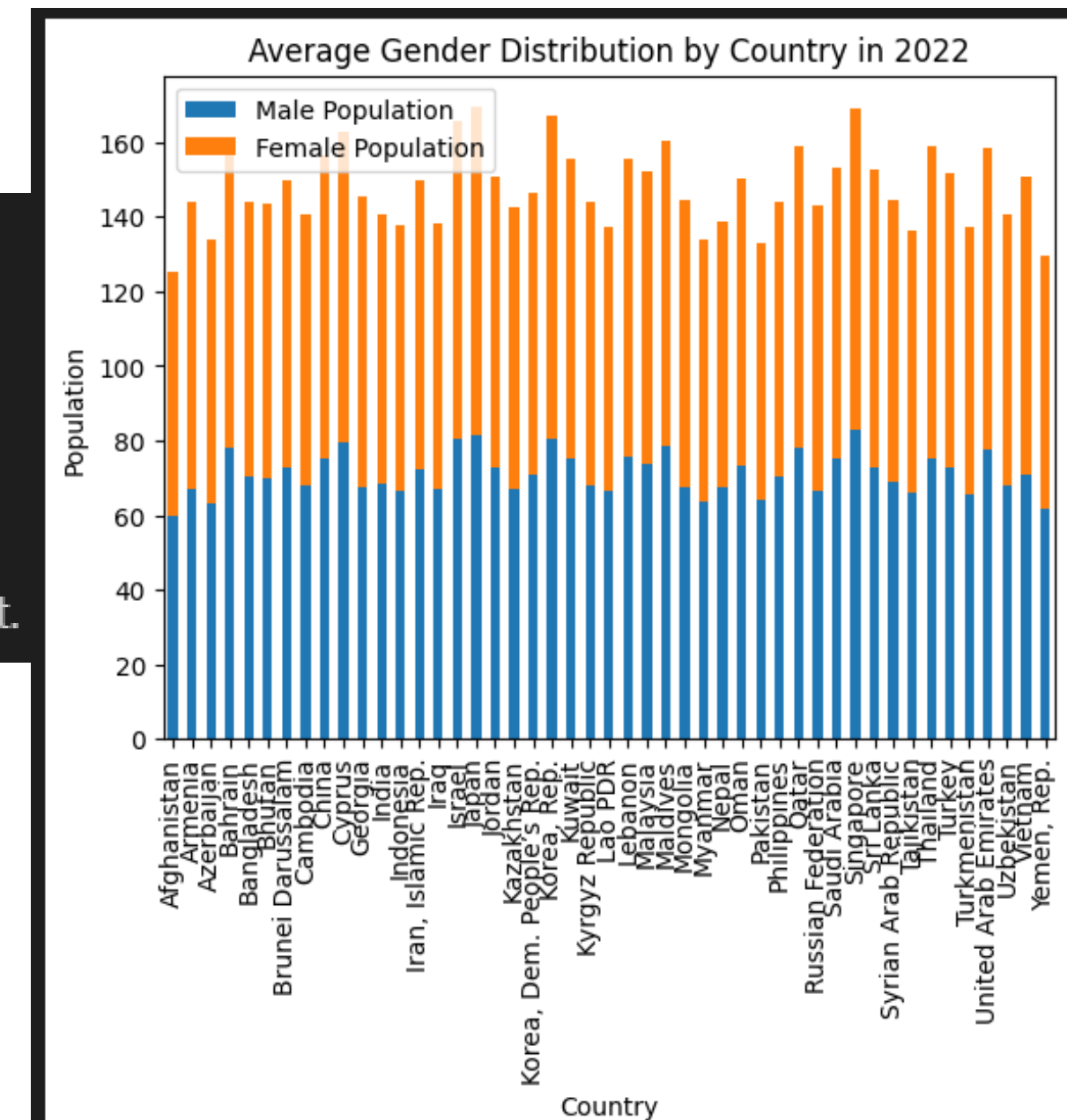
MAKE OWN QUESTION

- For more understanding about our data

Making your own question 04? (1.5 points)

A possible question is : What is top 5 country have the average percentage of students who complete primary school in each country?

Answering this question will: Show the top 5 averages of Primary completion rate for each continents.

How we answer this question: Calculate the average Primary completion rate for each continent.

	Avg.Primary completion rate	Country
42	109.053139	Turkmenistan
16	106.328373	Japan
10	102.607251	Georgia
26	102.517827	Maldives
45	102.214230	Vietnam

MAKE OWN QUESTION

- For more understanding about our data

Making your own question 05? (1.5 points)

A possible question is : What is top 5 country have the average highest unemployment rate ?

Answering this question will: Show the top 5 countries with the highest average unemployment rate

How we answer this question: Calculate the average unemployment rate for 5 continents.

	Avg.Unemployment rate	Country
17	14.601536	Jordan
10	14.276750	Georgia
46	12.364821	Yemen, Rep.
13	11.263286	Iran, Islamic Rep.
39	11.052964	Tajikistan

MODEL

MODEL

- Linear Regression
- Feedforward Neural Network (FNN)
- Recurrent Neural Network (RNN)

LINEAR REGRESSION

- Linear regression is a statistical method and a fundamental machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables. The core idea is to find the linear relationship that best describes the data.

FEEDFORWARD NEURAL NETWORK (FNN)

- A type of artificial neural network in which the information moves in only one direction forward from the input layer, through the hidden layers (if any), and finally to the output layer. There are no cycles or loops in the network.

FEEDFORWARD NEURAL NETWORK (FNN)

The three main types of layers in a typical FNN are:

- **Input Layer:** This layer receives the initial input data. Each neuron in this layer represents a feature or input variable.
- **Hidden Layers:** These layers come between the input and output layers, perform computations on the input data using weights and activation functions to produce an output that is passed to the next layer, and can have multiple hidden layers.
- **Output Layer:** The final output of the network.

FEEDFORWARD NEURAL NETWORK (FNN)
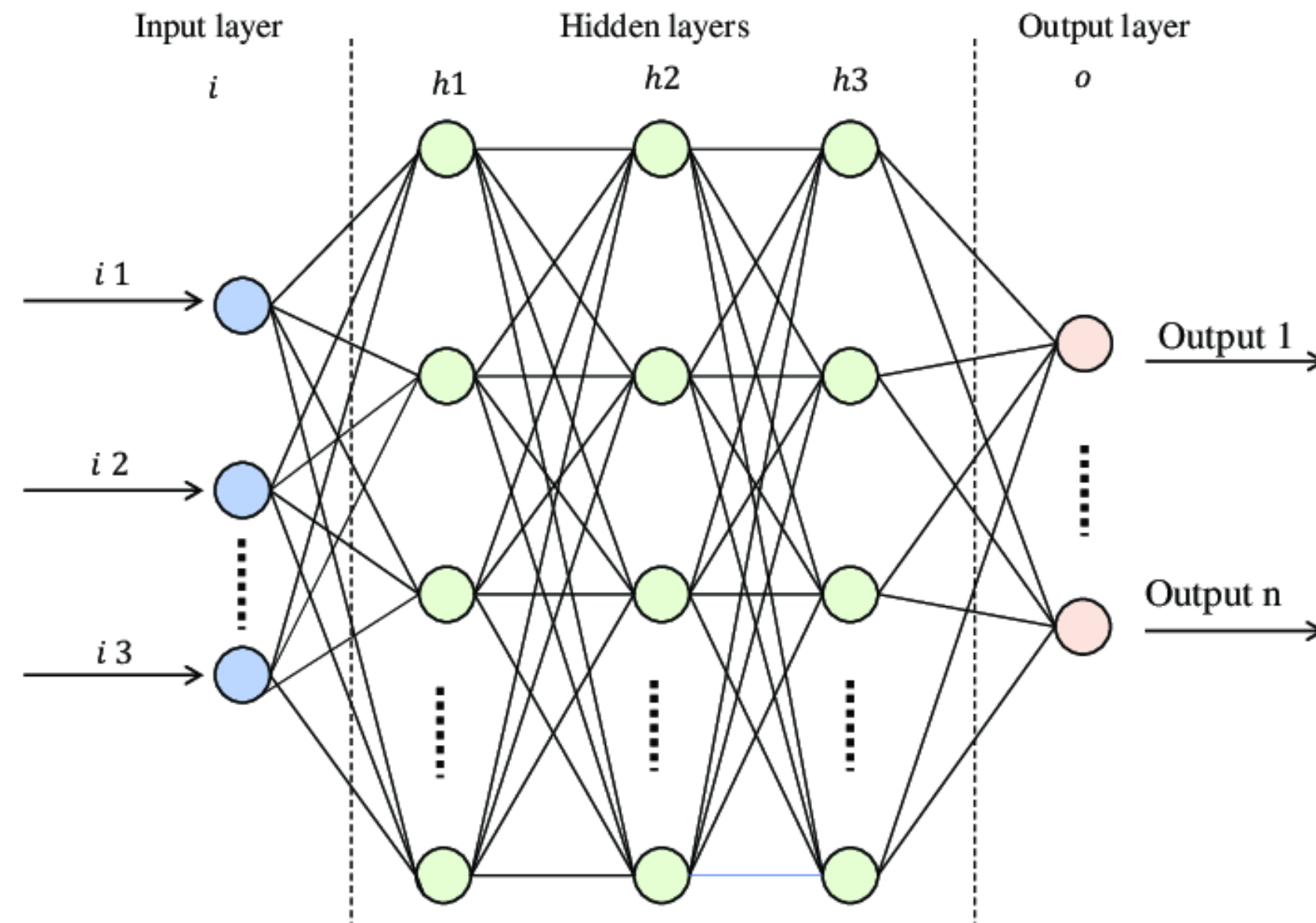
# RECURRENT NEURAL NETWORK (RNN)

- A Recurrent Neural Network (RNN) is a type of artificial neural network designed to process sequential and time-series data. RNNs are capable of handling sequence information by storing and utilizing information from previous time steps when computing information at the current time step.
- The structure of an RNN comprises "memory units" that can store information and access information from previous time steps. Each memory unit takes input from the current time step along with information from the previous time step to make predictions or generate results at the current time step.

RECURRENT NEURAL NETWORK (RNN)

The layers of a Recurrent Neural Network (RNN) typically include:

- **Input Layer**: This is the initial layer of the neural network, responsible for receiving input data and passing it forward to the subsequent layer. In the context of an RNN, the input layer often receives sequences of data, such as a sequence of words in a sentence.
- **Recurrent Layer**: This layer is pivotal in an RNN as it retains information from the past. It maintains a hidden state and takes input from the previous layer along with the previous hidden state to generate new outputs and update the hidden state.
- **Output Layer**: The final layer of the neural network, producing predictions or the model's output. In an RNN model, the output layer might be used to predict the next value in a sequence or generate predictions based on information from previous time steps.

RECURRENT NEURAL NETWORK (RNN)

Recurrent network

input layer

hidden layers: "deep" if > 1

output layer
(class/target)
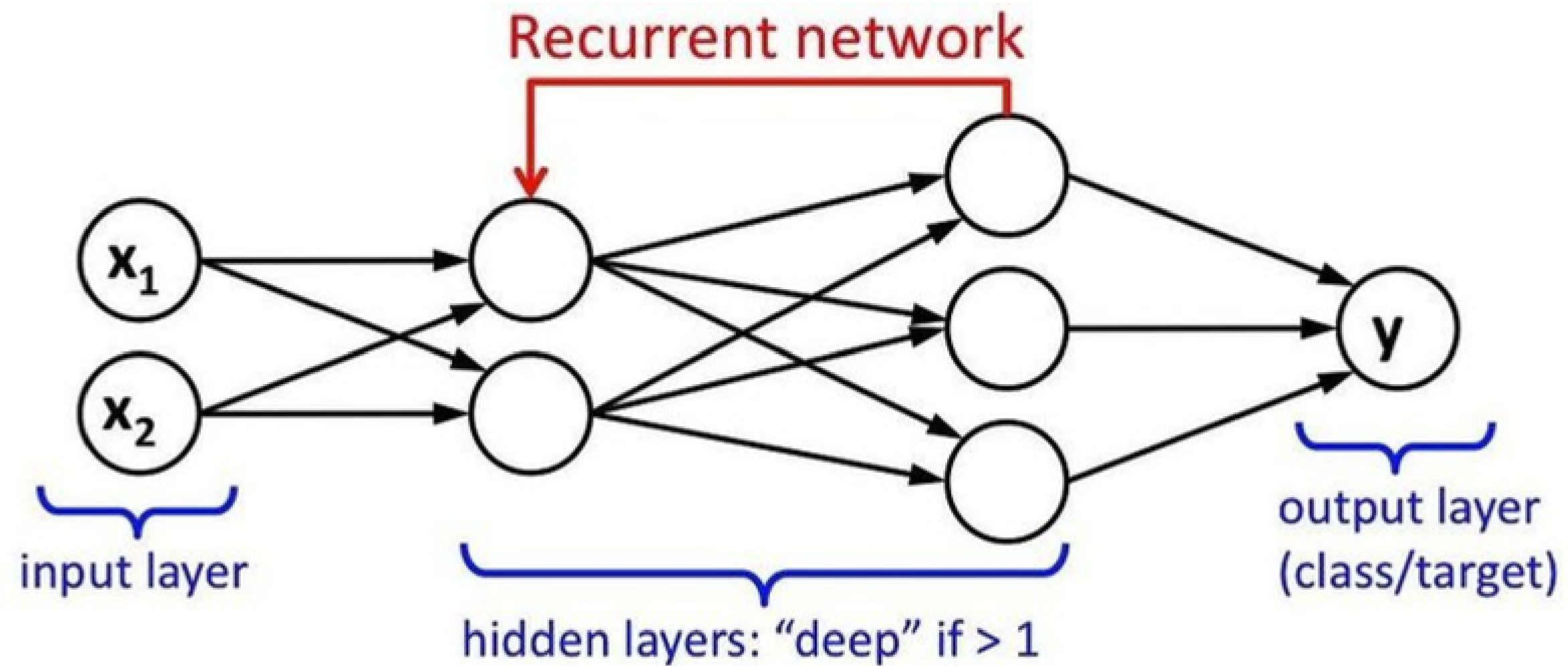
PREPROCESSING MODEL
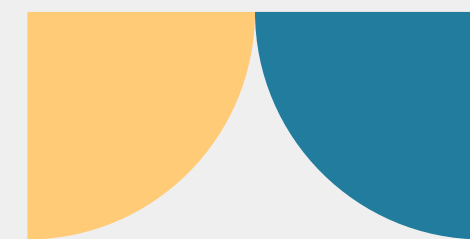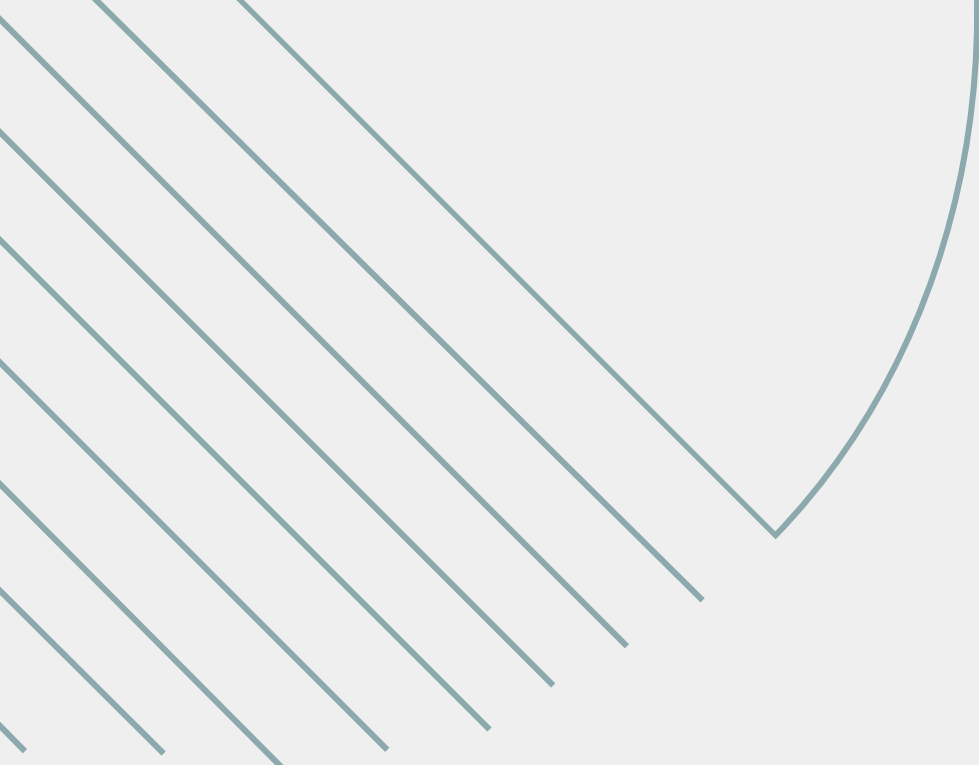
LINEAR REGRESSION

- Using label encoding to change the "Country" column to number for training
- Split data to train and test set for training
- Then apply StandardScaler for normalization across the data columns.

FEEDFORWARD NEURAL NETWORK (FNN)

- Split data to train, validate and test set
- Apply StandardScaler for normalization across the data columns.
- Concatenate the preprocessed and normalized columns into the feature dataset for training.
- Using label encoding to change the "Country" column to number for training
- Then concatenate the preprocessed and normalized columns into the feature dataset for training.

RECURRENT NEURAL NETWORK (RNN)

- Utilize One-Hot Encoding technique to transform the 'Country' column (names of countries) into integers since machine learning models often require numerical values as input data, not textual labels.
- Apply StandardScaler for normalization across the data columns.
- Concatenate the preprocessed and normalized columns into the feature dataset for training.
- Then, split the data into the training set, validation set, and test set.

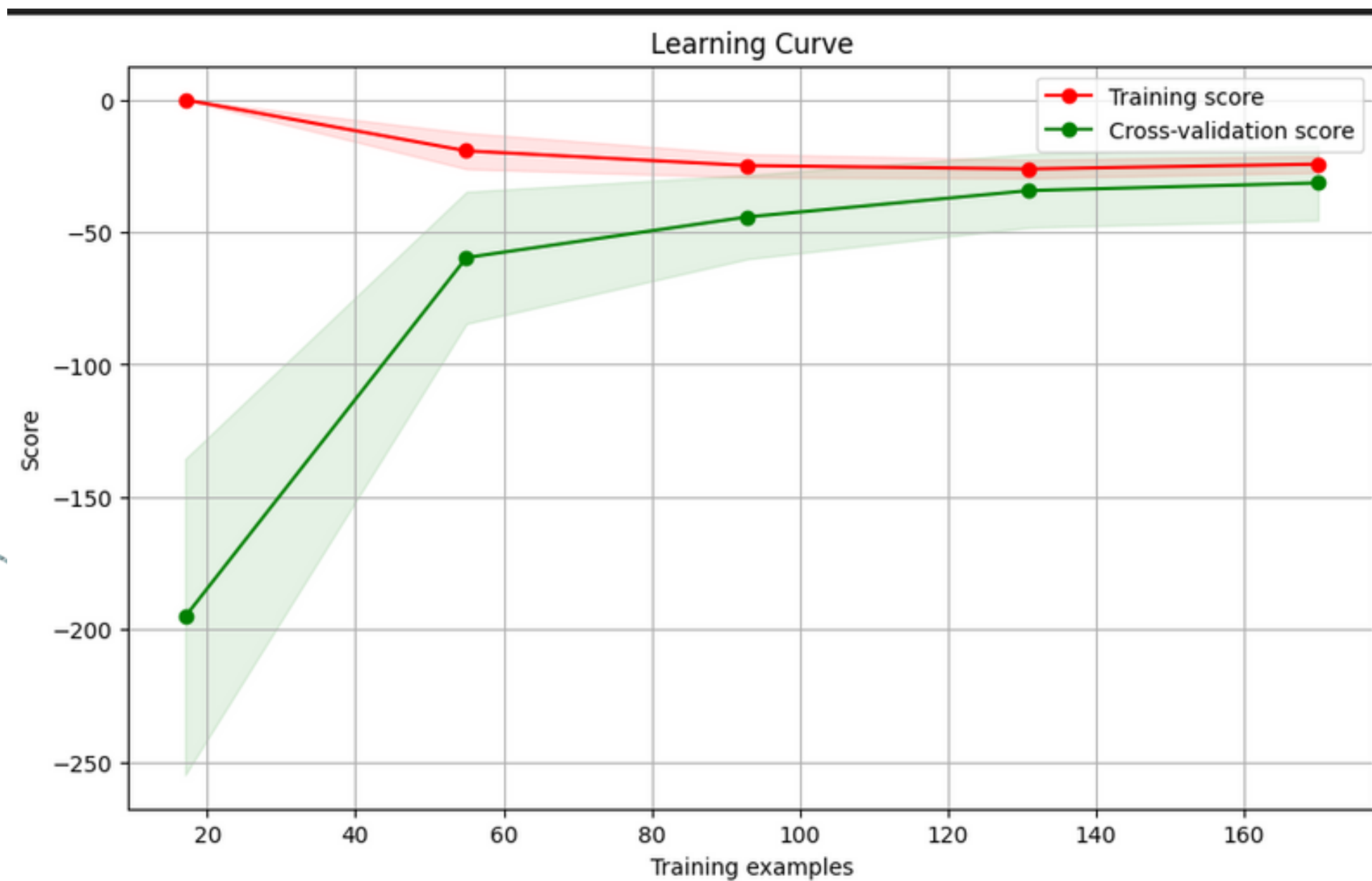# TRAIN & TEST

LINEAR REGRESSION

- Result train and test:

```
Linear Regression:
Mean Squared Error: 27.403293803258343
R-squared: 0.5566496633434844
```

Linear Regression model is providing a moderate level of performance on the given dataset. It's capturing a significant portion of the variance in the target variable, but there is room for improvement, especially if higher predictive accuracy is desired.

LINEAR REGRESSION

- Result:

Evaluate:
MSE when training for the first time, there is a high error with the training process and low results compared to the testing process, but later on the model is equal in terms of accuracy.

FEEDFORWARD NEURAL NETWORK (FNN)

```
13/13 [==============================] - 3s 39ms/step - loss: 9437.2207 - val_loss:
9185.7373
Epoch 2/200
13/13 [==============================] - 0s 13ms/step - loss: 8559.1221 - val_loss:
7406.4390
Epoch 3/200
13/13 [==============================] - 0s 14ms/step - loss: 5413.4375 - val_loss:
2450.2463
Epoch 4/200
```

```
...
Epoch 199/200
13/13 [==============================] - 0s 13ms/step - loss: 86.2247 - val_loss:
141.6702
Epoch 200/200
13/13 [==============================] - 0s 12ms/step - loss: 87.9455 - val_loss:
150.5807
```

FEEDFORWARD NEURAL NETWORK (FNN)

```
9/9 [==============================] - 0s 4ms/step - loss: 144.9937
Test Loss: 144.9937286376953
```

```
Mean Squared Error: 144.9937305678236
Mean Absolute Error: 11.750068057667125
```

FEEDFORWARD NEURAL NETWORK (FNN)
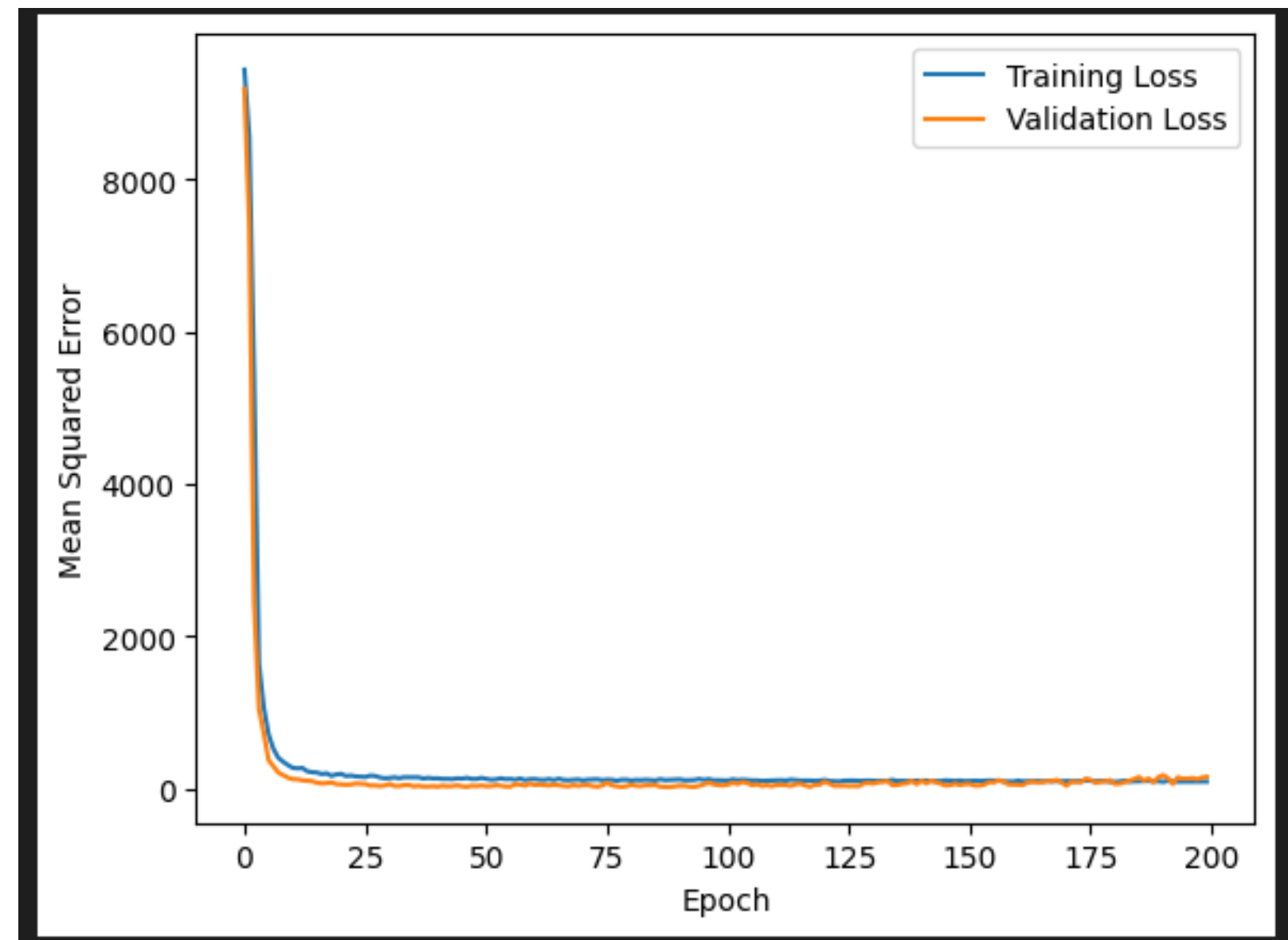
RECURRENT NEURAL NETWORK (RNN)

- kết quả train & test :


```
Epoch 2/200
3/3 [==============================] - 0s 25ms/step - loss: 8842.5996 - val_loss: 9194.2705
Epoch 3/200
3/3 [==============================] - 0s 25ms/step - loss: 8824.5225 - val_loss: 9175.8828
Epoch 4/200
3/3 [==============================] - 0s 39ms/step - loss: 8804.4102 - val_loss: 9154.6230
Epoch 5/200
3/3 [==============================] - 0s 27ms/step - loss: 8780.5684 - val_loss: 9123.7715
Epoch 6/200
3/3 [==============================] - 0s 25ms/step - loss: 8741.1416 - val_loss: 9040.7969
Epoch 7/200
3/3 [==============================] - 0s 23ms/step - loss: 8593.8916 - val_loss: 8543.3994
Epoch 8/200
3/3 [==============================] - 0s 21ms/step - loss: 7473.7729 - val_loss: 4518.2178
Epoch 9/200
3/3 [==============================] - 0s 23ms/step - loss: 3521.8428 - val_loss: 697.8221
Epoch 10/200
3/3 [==============================] - 0s 24ms/step - loss: 1005.6748 - val_loss: 2757.7976
Epoch 11/200
3/3 [==============================] - 0s 30ms/step - loss: 2857.9089 - val_loss: 2991.4316
Epoch 12/200
3/3 [==============================] - 0s 26ms/step - loss: 2080.3291 - val_loss: 428.0960
Epoch 13/200
3/3 [==============================] - 0s 25ms/step - loss: 741.3483 - val_loss: 1204.2307
...
Epoch 200/200
3/3 [==============================] - 0s 21ms/step - loss: 107.7399 - val_loss: 65.4188
2/2 [==============================] - 0s 5ms/step - loss: 69.1145
2/2 [==============================] - 0s 3ms/step
```

RECURRENT NEURAL NETWORK (RNN)

- Visualize :
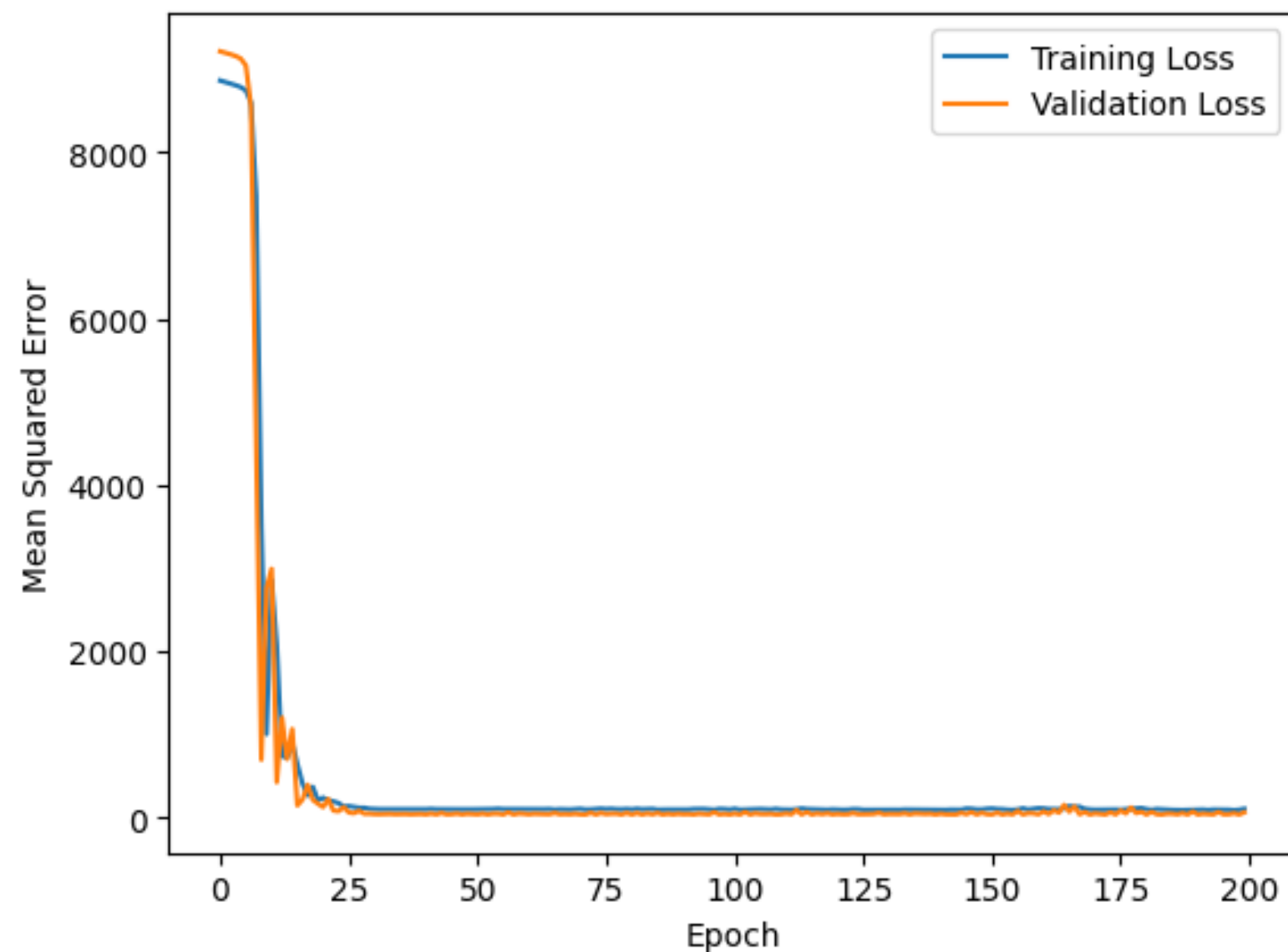
- Evaluate :

- Both the training error curve (in blue) and the validation error curve (in orange) exhibit a substantial decrease as the number of iterations increases, indicating an effectively learning model.
- Around 25 iterations, both curves stabilize, suggesting minimal reduction in error with increasing iterations, indicating convergence.
- The mean squared error (MSE) of the model is approximately 500 at the 200th iteration, which could be considered relatively low depending on the problem and dataset.

COMPARISON BETWEEN MODELS

- Linear Regression is a basic model, is the first model to think about when there's some problem, but the there's many limited, but in our project, it has the small MSE, which mean linear regression is pretty suitable with our project
- FNN is artificial neural network, in our project, the MSE is biggest
- CNN is artificial neural network, in our project, the MSE is smallest

REFERENCE

- https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable
- https://www.youtube.com/watch?v=6AK7qC2XGHY&t=12s
- https://nttuan8.com/bai-13-recurrent-neural-network/
- https://www.youtube.com/watch?v=nwD5U2WxTdk&list=PLuhqtP7jdD8AFocJuxC6_Zz0HepAWL9cF

THANK YOU

Cảm ơn thầy đã lăng nghe