Tasks

Task 1: Preprocess the data

After you've run the script `generate_data.py` you should have csv files in the directory `data/generated`. You should have one file for each country. Each file should have data that looks like this:

Year	Consumption_Coal	Consumption_NaturalGas	Consumption_Neuclear+renewables	Consumption_Petroleum	Consumption_Total	Production_
1980	0	0.0096324	0.005505642	0.043228106	0.058366148	0
1981	0	0.0092136	0.005540091	0.043109997	0.057863688	0
1982	0	0.01047	0.006272403	0.045265496	0.062007899	0
1983	0	0.012564	0.006312011	0.047420996	0.066297007	0

The first column is the year. The rest of the columns are the features. The features are the consumption and production of each energy source. Some files will have `--` entries in years that it didn't exist during. However, if a country exists but did not have any production or consumption of a certain energy source, then the entry will be `0`.

You'll need to do the following:

- In the root directory of this project, create a script called `preprocess_data.py`.
- This script should read each csv files in the directory `data/generated` and preprocess it and write it to the directiory `data/preprocessed`. The preprocessed data should be in the same table structure as the original data.
- For countries that were separated or merged (like sudan and south sudan), you need to merge the parent country's old data with the new country's missing data. This is not accurate however it's a close enough estimate. (LOL the amount of politics you'll need to read up on is massive to do that task or you can just remove those years from the dataset but I know they won't like it)
- You'll need to convert the data from quadrillion british thermal units (quad btu) to terawatt hours (twh). 1 quad btu = 293.071 twh. You can use the following function to convert the data:

```
def quad_btu_to_twh(quad_btu: float) -> float:
    return quad_btu * 293.071
```

- You still need to normalize the data. I think z-score normalization is best here.

Task 2: Create Models

Now that you've preprocessed the data, you can now use models to forecast the data. For each country, we want to forecast the consumption and production of each energy source for the next x years (cause still i don't know how many years some models are able to forecast). The models you'll be using are:

- Using the tensorflow library:
  - RNN
  - LSTM
  - GRU

The ARIMA and other traditional models will be up to nour.

Task 3: Evaluate Models

You'll need to evaluate the models using the following metrics:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

You'll need to create a script called `evaluate_models.py` that will evaluate each model against each country. You'll need to write the results to a csv file in the directory `data/evaluation`. The csv file should look like this just for example:

Country	Model	MAE	MSE	RMSE	MAPE
USA	RNN	0.1	0.2	0.3	0.4
USA	LSTM	0.1	0.2	0.3	0.4
USA	GRU	0.1	0.2	0.3	0.4
USA	ARIMA	0.1	0.2	0.3	0.4

Task 4: Visualize Results

You'll need to create a script called `visualize_results.py` that will visualize the results of the evaluation. You'll need to create a directory called `data/visualization` and save the visualizations there.

End Notes

Some of those task require some collaborative work since they are related to each other. You'll do the preprocessing purely alone and share with nour some modeling and evaluation work and you'll share analysis and visualization tasks with me.