**Measure Energy Consumption**

# *Phase 3:* Assignment Notebook Submission

**Energy Consumption:** Refers To ALL The Energy Used In Day To Day Life. There are many factors at play when it comes to measuring the energy consumption of a country, mainly the consumption difference between a countries industries and its population.

**Important to know: Countries population vs industries:**

* The energy consumption of a population is often measured using Energy intensity per capita, which we looked at a few sections above. In that section we saw that more developed countries tend to have a worse Energy-intensity per Capita, as there population is able to spend more on amenities like technology, appliances, and transportation. We also saw that the countries with the smallest energy intensity per capita were all developing countries, there populations arent able to afford many of the amenities available to more developed nations.
* The energy consumption of industries is often measured using Energy intensity of GDP, which we also saw a few sections above. In that section we saw that more developed countries tend to have a smaller energy intensity of GDP, while devolping countries tend to have a larger energy intensity of GDP. The main reason for this is that more developed countries can afford less energy intensive technologies and ways of production, unlike developing countries which cant, therfore leaving them stuck with more energy intensive means of production.

**Important to know: Countries population vs industries**

The energy consumption of a population is often measured using Energy intensity per capita, which we looked at a few sections above. In that section we saw that more developed countries tend to have a worse Energy-intensity per Capita, as there population is able to spend more on amenities like technology, appliances, and transportation. We also saw that the countries with the smallest energy intensity per capita were all developing countries, there populations arent able to afford many of the amenities available to more developed nations. The energy consumption of industries is often measured using Energy intensity of GDP, which we also saw a few sections above. In that section we saw that more developed countries tend to have a smaller energy intensity of GDP, while devolping countries tend to have a larger energy intensity of GDP. The main reason for this is that more developed countries can afford less energy intensive technologies and ways of production, unlike developing countries which cant, therfore leaving them stuck with more energy intensive means of production.

**Table of Contents**

* Cleaning Data
* Data Structure
* Analysis
* Conclusion

**Categorization of different countries:**

Categorizing countries into distinct groups based on its development is somewhat complex, and when it comes to classifying this, there is no single way (either grounded in theory or based on an objective benchmark) that is gererally accepted. The UN model, or there World Economic Situation and Prospects (WESP) report to be exact classifies every country into one of three broad categories: Developed Economies, Economy in transition, and Developing economy. This is the categorization Ill be using throughout this notebook.

**Warning:**

I am not a climate scientist, some things may be inacurate. This is simply just a study on a subject im interested in, allowing me to go deeper into the subject while at the same time imporving my graphing skills. All my sources are at the bottom of the notebook.

**Column descriptions:**

* **Country** - Country in question
* **Energy\_type** - Type of energy source
* **Year** - Year the data was recorded
* **Energy\_consumption** - Amount of Consumption for the specific energy source, measured (quad Btu)
* **Energy\_production** - Amount of Production for the specific energy source, measured (quad Btu)
* **GDP** - Countries GDP at purchasing power parities, measured (Billion 2015$ PPP)
* **Population** - Population of specific Country, measured (Mperson)
* **Energy\_intensity\_per\_capita** - Energy intensity is a measure of the energy inefficiency of an economy. It is calculated as units of energy per unit of capita (capita = individual person), measured (MMBtu/person)
* **Energy\_intensity\_by\_GDP**- Energy intensity is a measure of the energy inefficiency of an economy. It is calculated as units of energy per unit of GDP, measred (1000 Btu/2015$ GDP PPP)
* **CO2\_emission** - The amount of C02 emitted, measured (MMtonnes CO2)

*# Analysis Tools*

**import** numpy **as** np

**import** pandas **as** pd

**from** scipy **import** stats

**from** scipy.stats **import** norm

​

*# Plotting Tools*

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

**import** plotly.express **as** px

**import** plotly.graph\_objects **as** go

**from** plotly.subplots **import** make\_subplots

​

*# Extra Plotting Tools Required for Bar Chart Race*

**import** matplotlib.ticker **as** ticker

**import** matplotlib.animation **as** animation

**from** IPython.display **import** HTML

*# Plot Design Settings*

sns.set\_style("darkgrid", {"axes.facecolor": "#eff2f5", 'grid.color': '#c0ccd8', 'patch.edgecolor': '#B0B0B0', 'font.sans-serif': 'Verdana'})

sns.set\_palette('Dark2\_r')

plt.rc('font', size**=**19)

plt.rc('axes', titlesize**=**25)

plt.rc('axes', labelsize**=**20)

plt.rc('xtick', labelsize**=**17)

plt.rc('ytick', labelsize**=**17)

plt.rc('figure', titlesize**=**24)

​

*# Other Tools*

**from** sklearn.preprocessing **import** OneHotEncoder

​

*# Mute warnings*

**import** warnings

warnings.filterwarnings('ignore')

## ****Cleaning Data****

df **=** pd.read\_csv("../input/c02-emission-by-countrys-grouth-and-population/energy.csv")

df.shape

df.head(6)

*# Removing extra index column*

df **=** df.drop(['Unnamed: 0'], axis**=**1)

### Renaming columns for simplicity

Some column names are quite long, im just going to shorten them to abreviations

df.rename(columns**=**{'Energy\_type' : 'e\_type', 'Energy\_consumption' : 'e\_con', 'Energy\_production' : 'e\_prod'

, 'Energy\_intensity\_per\_capita' : 'ei\_capita', 'Energy\_intensity\_by\_GDP' : 'ei\_gdp'}, inplace**=True**)

df.head()

Great, much less complicated to write now

### Renaming e\_type value names

Some of the values for the e\_type column are very long, im going to shorten them

df['e\_type'] **=** df['e\_type'].astype('category')

​

df['e\_type'] **=** df['e\_type'].cat.rename\_categories({'all\_energy\_types': 'all', 'natural\_gas': 'nat\_gas','petroleum\_n\_other\_liquids': 'pet/oth',

'renewables\_n\_other': 'ren/oth'})

​

df['e\_type'] **=** df['e\_type'].astype('object')

### Data Types

Converting Year column to datetime dtype

df['Year'] **=** df['Year'].astype('object')

df.info()

### Columns Stats

df.describe(include**=**'all')

### Features Unique Values

*# Number of unique values in each variable*

**for** var **in** df:

print(f'{var}: {df[var].nunique()}')

### Duplicates

print('Number of Duplicates: {}'.format(len(df[df.duplicated()])))

Cool, no duplicate values

### Missing Values

There is quite a bit of data thats missing in this dataset. For example, some of the missing data is because the given country no longer exists, leaving only NaN values before/after its creation/collapse, theres is also a lot of data missing not at random (MNAR), as well as data missing at random (MAR). I will be dealing with each of these appropriatlly in this section

**for** var **in** df:

print(f'{var}: {df[var].isnull().sum()}')

*#Missing data as white lines*

**import** missingno **as** msno

msno.matrix(df,color**=**(0.3,0.36,0.44))

*# Function ill be using to drop selected coutries*

**def** to\_drop(list):

**for** country **in** list:

value **=** df[df['Country']**==**country].index

df.drop(labels**=**value, axis**=**0, inplace**=True**)

**Taking care of countries that no longer exist or formed within the time period**

Quite a bit of the Countries/Territories in this dataset no longer exist or have become another country, leaving only NaN values for the years they didnt exist. Im going to drop all the rows of these years. Below are some examples of some of these countries.

df[df['Country']**==**'Former U.S.S.R.']

df[df['Country']**==**'South Sudan']

For all the years that each of the countries didnt exist, a set of NaN values remain for its columns. Luckily, theres a tool we can use to simply drop all rows with a certain amount of NaN values in them, which is dropna(). Im going to drop all the rows of the countries with more than 3 values set to NaN as this will take care of the years these countries didnt exist as well as rows that have to little information.