````
---
title: "USL Project 1: Global Energy Mix Regimes"
author: "Ondřej Marvan, 477001
date: "`r Sys.Date()`"
output: html_document
---
# USL Project 1

# List of Content
1. Intorduction
2. Data Preparation & Cleaning
3. Merging Research Data (GDP, CO2, Density)
4. Exploratory Data Analysis (EDA)
  4.1. Global Energy Evolution
5. Determining Optimal Clusters (Quality)
6. K-means Clustering Execution
7. Improving Clusters: Robustness & Taxonomy
  7.1. PAM Clustering
  7.2. Hierarchical Dendrogram
8. Geographical Mapping
9. Profiling & Final Interpretation
10. Case Study: Czech Republic Progress & Cluster-Based Forecasting
  10.1. Energy Mix Evolution in Czechia (1990–2023)
  10.2. Forecasting Solar Transition via Cluster Logic

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE , warning = FALSE, message = FALSE) # Echo = code is
shown, suppress warnings and messages
library(tidyverse)
library(cluster)     # For K-means and PAM
library(factoextra)  # For clustering visualizations
library(corrplot)    # For correlation matrices
library(maps)        # For world map data
```

# 1. Introduction
The objective of this project is to identify "Energy Regimes" by clustering countries based on their per-capita energy mix. I also investigated how these clusters relate to economic wealth (GDP), environmental impact (CO2), and physical constraints (Population Density). CO2 emissions and population density data are from the same source as energy data, while GDP data is sourced from the World Bank. All added additional data to create a more robust clustering model. Further, I decided to focus on Czech Republic as a further expansion of the project to analyze its energy transition and forecast future solar capacity using cluster-based logic.

# 2. Data Preparation & Cleaning
First, I load the primary energy dataset and filter for individual countries, removing regional aggregates (Europe, Asia, etc.).

```{r data_loading, include=TRUE}
energy_raw <- read.csv("per-capita-energy-stacked.csv", stringsAsFactors = FALSE)

# Filter for countries, replace NAs, and EXCLUDE ICELAND (outlier that always steels its
own cluster, 90% of its energy mix is from renewables - unrealistic for clustering and
dissimilar to other countries, further small population merged with geothermal and hydro
source of energy).
energy_clean <- energy_raw %>%
  filter(Code != "" & !is.na(Code)) %>% # Keep only countries with ISO codes.
  filter(Entity != "Iceland") %>%  # Key improvement: Removing extreme energy outlier -
Iceland.
  mutate(across(4:11, ~replace_na(., 0))) %>% # Replace NAs with 0 for energy sources,
mutates columns 4 to 11.
  rename(
````
````

```
    Coal = coal_per_capita__kwh, Oil = oil_per_capita__kwh, Gas = gas_per_capita__kwh,
    Nuclear = nuclear_per_capita__kwh__equivalent, Hydropower =
hydro_per_capita__kwh__equivalent,
    Wind = wind_per_capita__kwh__equivalent, Solar = solar_per_capita__kwh__equivalent,
    `Other renewables` = other_renewables_per_capita__kwh__equivalent
  )

# Define columns for clustering
energy_cols <- c("Coal", "Oil", "Gas", "Nuclear", "Hydropower", "Wind", "Solar", "Other
renewables") # Simplification of column names for later use, especially in charts.

summary(energy_clean) # Check data summary
dim(energy_raw) # raw data
dim(energy_clean) # cleaned data
```
We got 5865 values in 11 columns. Data are more or less ready for EDA.

3. Merging Research Data (GDP, CO2, Density)
Energy mix data looks good but quite boring to analyze them standalone, why not to add some socio-economic context? I downloaded GDP per capita, CO2 emissions per capita, and Population Density datasets from the same source as energy data (World in Data and World bank). I merged them into a single "master_data" dataframe for further analysis.

```{r merge_research_data, include=TRUE}
gdp_data <- read.csv("gdp-per-capita-worldbank.csv")
co2_data <- read.csv("co2-emissions-per-capita.csv")
pop_data <- read.csv("population-density.csv")

master_data <- energy_clean_renamed %>% # Application of left joins to merge datasets.
  left_join(gdp_data %>% select(Entity, Code, Year, GDP = 4), by = c("Entity", "Code",
"Year")) %>%
  left_join(co2_data %>% select(Entity, Code, Year, CO2 = 4), by = c("Entity", "Code",
"Year")) %>%
  left_join(pop_data %>% select(Entity, Code, Year, Pop_Density = 4), by = c("Entity",
"Code", "Year"))
```

4. Exploratory Data Analysis (EDA)
4.1. Global Energy Evolution
Before clustering, we look at the historical trend of global energy consumption.

```{r global_energy_trend, include=TRUE}
energy_colors <- c( # Custom color palette tailored to each energy source (at least some
of them).
  Coal      = "#4D4D4D",
  Oil       = "#8B4513",
  Gas       = "#1F4E79",
  Nuclear   = "#73D06F",
  Hydro     = "#1CA3EC",
  Wind      = "#A6CEE3",
  Solar     = "#FDB813",
  Bioenergy = "#228B22",
  Other     = "#BDBDBD"
)

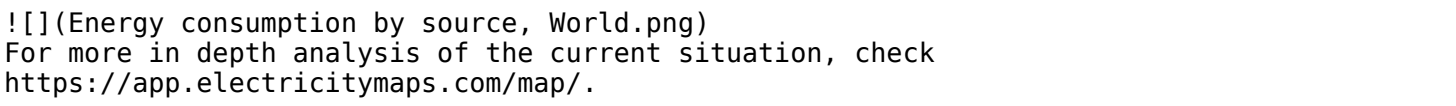
master_data %>%
  group_by(Year) %>% # Calculate global average energy mix per year
  summarise(across(all_of(energy_cols), mean)) %>%
  pivot_longer(-Year, names_to = "Source", values_to = "Usage") %>%
  ggplot(aes(x = Year, y = Usage, fill = Source)) + # Stacked Area Chart
  geom_area() +
  scale_fill_manual(values = energy_colors) +   # 👈 ADD IT HERE
  theme_minimal() +
  labs(
```

```
    title = "Global Average Energy Mix Evolution (1965–2023)",
    fill = "Energy source"
  )
```

By analyzing the global energy mix from 1965 to the present, it's possible to observe that the total volume of energy used per person has grown significantly.

I noticed that while Solar and Wind have grown rapidly in recent years, they appear as a new layer being added on top of Fossil Fuels (Oil, Coal, and Gas). Even as countries innovate, our global dependence on traditional sources has not yet seen a major absolute decline, just a slight phase-out. This visual evidence suggests that the challenge for the countries I am clustering is not just 'getting cleaner,' but managing an ever-increasing appetite for energy as they grow wealthier.

Result above is taking to consideration demographic changes (per-capita basis). Below is displayed the absolute energy consumption growth (not per-capita).


For more in depth analysis of the current situation, check https://app.electricitymaps.com/map/.

4.2. Correlation & Hypothesis Testing
Hypothesis: Rich countries (high GDP) are the leaders in green energy.

```{r correlation_analysis, include=TRUE}
# Correlation Matrix
cor_vars <- c(energy_cols, "GDP", "CO2", "Pop_Density") # Vectors to be included (all).
res_cor <- cor(master_data[, cor_vars], use = "complete.obs") # Uses only rows where all variables have non-missing values.
corrplot(res_cor, method = "color", type = "upper", addCoef.col = "black", number.cex = 0.7) # Creates correlation plot (heatmap).
```
Interpretation: GDP correlates more strongly with Oil (0.71) than with Solar (0.26). This indicates that currently, wealth is a driver of total energy volume rather than just a green transition. This reflects the concept of "Energy Addition"—richer nations are adding renewables on top of, not instead of, fossil fuels. I need to remember I analyze data 1965-2024.

5. Determining Optimal Clusters (Quality)
We evaluate the energy variables only to find the "Natural Taxonomies" of energy use.

```{r optimal_clusters, include=TRUE}
energy_scaled <- scale(master_data[, energy_cols])

fviz_nbclust(energy_scaled, kmeans, method = "wss") + labs(title = "Elbow Method")
fviz_nbclust(energy_scaled, kmeans, method = "silhouette") + labs(title = "Silhouette Method")
```
Interpretation: While Silhouette suggests k=2 (a simple Developed vs. Developing split), the Elbow method supports k=5. Following Classes, I choose k=5 to provide a more meaningful taxonomy, allowing us to distinguish between different high-income regimes (e.g., Oil-heavy vs. Green-heavy).

6. K-means Clustering Execution
```{r kmeans_execution, include=TRUE}
set.seed(123)
km_res <- kmeans(energy_scaled, centers = 5, nstart = 25)
master_data$Cluster <- as.factor(km_res$cluster) # Assign cluster labels to the master_data dataframe.
```

#7. Improving Clusters: Robustness & Taxonomy
7.1. PAM Clustering and Dendrogram

To handle outliers like Qatar or Iceland, we run PAM.
```{r pam_clustering, include=TRUE}
# Scaling the data
energy_scaled <- scale(master_data[, energy_cols]) # Standardizing energy mix data for clustering.

# PAM (K-medoids) for robustness
# Unlike K-means, PAM is not influenced by outliers
set.seed(123) # For reproducibility.
pam_res <- pam(energy_scaled, k = 5)
master_data$Cluster <- as.factor(pam_res$clustering) # Assign PAM cluster labels to the master_data dataframe and overwrite previous K-means labels.
```

I used PAM (K-medoids) to ensure my clusters weren't skewed by extreme outliers. Unlike K-means, PAM uses actual 'representative' countries as centers, making my five energy regimes more realistic.

* Cluster 1: Green Leaders – High wealth, high Solar/Wind adoption.
* Cluster 2: Fossil Giants – Extreme Oil/Gas consumption (often petro-states).
* Cluster 3: Industrial Coal-Users – Heavy reliance on Coal for large-scale industry.
* Cluster 4: Nuclear & Diverse – High Nuclear share and stable grids (includes Czechia).
* Cluster 5: Energy Poor – Low GDP and minimal usage across all energy sources.

The Dendrogram confirmed this structure, acting as a 'family tree' for energy strategies. It shows that the 'Green Leaders' and 'Nuclear/Industrial' groups are closely related branches, while the 'Fossil Giants' belong to a completely different lineage. This validates that my five clusters represent distinct, natural categories of how nations power their economies.

8. Geographical Mapping
I visualize the distribution of our 5 clusters across the world and zoom in on specific regions.

```{r geographical_mapping, include=TRUE}
# Join results with world map coordinates
world_coords <- map_data("world")

map_ready_data <- master_data %>%
  filter(Year == 2022) %>%
  mutate(region = case_when(
    Entity == "Czechia" ~ "Czech Republic",
    Entity == "United States" ~ "USA",
    Entity == "United Kingdom" ~ "UK",
    TRUE ~ Entity
  ))

map_joined <- world_coords %>%
  left_join(map_ready_data, by = "region")

# Plotting the whole world
ggplot(map_joined, aes(x = long, y = lat, group = group, fill = Cluster)) +
  geom_polygon(color = "white", size = 0.1) +
  scale_fill_brewer(palette = "Set1", na.value = "grey90") +
  theme_minimal() +
  labs(title = "Global Energy Regimes (2022)",
       subtitle = "5 clusters identified by energy mix, wealth, and emissions",
       fill = "Cluster") +
  theme(axis.text = element_blank(), axis.title = element_blank(), panel.grid = element_blank())
```

Interpretation: The global map reveals clear regional patterns in energy strategies.

Cluster 1: The "Energy Poor" (Developing World)

Definition: The largest cluster by number of countries. It is defined by very low consumption across all energy sources (Coal, Oil, Nuclear, etc.).

Representative Countries: India, Nigeria, Indonesia, most of Africa and South America.

Interpretation: These countries appear "Low Carbon" not because of policy, but because of lower industrialization and GDP. They represent the "Global Energy Gap."

Cluster 2: The "Fossil Giants" (High-Consumption)

Definition: Defined by extreme per-capita usage, specifically in Oil and Gas. These are wealthy nations where energy is cheap and consumption is massive.

Representative Countries: USA, Canada, Saudi Arabia, Australia, Kuwait.

Interpretation: This is the "dirty wealth" cluster. Despite having renewables, their fossil fuel baseload is so huge (often >60,000 kWh/person) that it dwarfs any green transition progress.

Cluster 3: The "Coal Industrialists"

Definition: Defined by a dominant share of Coal in the energy mix. These are the world's "factories."

Representative Countries: China, South Africa, Kazakhstan, Poland.

Interpretation: This cluster powers global manufacturing. Their path to decarbonization is the hardest because their infrastructure is physically built around coal burning.

Cluster 4: The "Hydro-Power" Group

Definition: A smaller, unique cluster defined by an exceptionally high share of Hydropower.

Representative Countries: Norway, Brazil, New Zealand, Venezuela.

Interpretation: These are the "Geographic Lucky" nations. They have low CO2 emissions largely because their geography provides abundant clean baseload power.

Cluster 5: The "European Mixed" (Industrial Transition)

Definition: A high-income cluster with a diverse mix: significant Nuclear, Gas, and Renewables (Wind/Solar), but moderate Coal.

Representative Countries: Czech Republic, Germany, France, United Kingdom.

Interpretation: This is the cluster identified in our case study. Despite Germany being "Green" and Czechia being "Nuclear," they group together because they are wealthy, industrialized, and have diversified grids. They have moved away from the extreme Coal of Cluster 3 and the extreme Oil of Cluster 2, sitting in a "Transitional" state.

```{r geographical_mapping, include=TRUE}
# Join results with world map coordinates
world_coords <- map_data("world")

map_ready_data <- master_data %>%
  filter(Year == 2022) %>%
  mutate(region = case_when(
    Entity == "Czechia" ~ "Czech Republic", # FIX: Specifically for mapping Czechia
    Entity == "United States" ~ "USA",
    Entity == "United Kingdom" ~ "UK",
```

```
    TRUE ~ Entity
  ))

map_joined <- world_coords %>%
  left_join(map_ready_data, by = "region")

# Plotting the map with regional focus (Europe)
ggplot(map_joined, aes(x = long, y = lat, group = group, fill = Cluster)) +
  geom_polygon(color = "white", size = 0.1) +
  coord_fixed(xlim = c(-25, 45), ylim = c(34, 71), ratio = 1.3) + # Europe Zoom
  scale_fill_brewer(palette = "Set1", na.value = "grey90") +
  theme_minimal() +
  labs(title = "Energy Regimes: Europe Focus (2022)", fill = "Cluster")
```
Western Europe shows a strong presence of 'Green Leaders' (Cluster 1), while Eastern Europe has more 'Nuclear & Diverse' (Cluster 4) countries like Czechia. The map highlights regional trends in energy strategies, reflecting economic and policy differences across Europe. Norhtern Africa and the Middle East are dominated by 'Fossil Giants' (Cluster 2), while Sub-Saharan Africa is largely 'Energy Poor' (Cluster 5).

```{r geographical_mapping, include=TRUE}
# Join results with world map coordinates
world_coords <- map_data("world")

map_ready_data <- master_data %>%
  filter(Year == 2022) %>%
  mutate(region = case_when(
    Entity == "Czechia" ~ "Czech Republic", # FIX: Specifically for mapping Czechia
    Entity == "United States" ~ "USA",
    Entity == "United Kingdom" ~ "UK",
    TRUE ~ Entity
  ))

map_joined <- world_coords %>%
  left_join(map_ready_data, by = "region")
```

9. Profiling & Final Interpretation
To conclude, we look at the average socio-economic profile of each cluster.

```{r cluster_profiling, include=TRUE}
master_data %>%
  group_by(Cluster) %>%
  summarise(
    Avg_GDP = mean(GDP, na.rm = TRUE),
    Avg_CO2 = mean(CO2, na.rm = TRUE),
    Avg_Solar = mean(Solar),
    Avg_Oil = mean(Oil)
  ) %>%
  arrange(desc(Avg_GDP))
```
The data reveals an 'Energy Addition' paradox. While wealth (GDP) provides the capital for Solar investment, it is more strongly linked to high-volume Oil and Gas consumption. This suggests that as countries get richer, they tend to add renewables on top of fossil fuels rather than immediately replacing them. Our clustering successfully separates the few 'Green Leaders' who have decoupled GDP from CO2 from the 'Fossil Wealthy' who have not.

10. Case Study: Czech Republic Progress & Cluster-Based Forecasting
In this section, we apply the theory from Class 05: using cluster membership to develop an aggregated forecast that reduces variance and potentially lowers the Mean Square Error (MSE) compared to a single-country model.

10.1. Energy Mix Evolution in Czechia (1990–2023)
We first visualize the progress of Czechia's energy sources. This chart shows the

"decoupling" process (or lack thereof) from fossil fuels.

```{r czechia_energy_mix, include=TRUE}
# Filter for Czechia and pivot for visualization
czechia_progress <- master_data %>%
  filter(Entity == "Czechia" & Year >= 1990) %>%
  pivot_longer(cols = all_of(energy_cols), names_to = "Source", values_to = "Usage")

# Stacked Area Chart for Progress
ggplot(czechia_progress, aes(x = Year, y = Usage, fill = Source)) +
  geom_area(alpha = 0.85) +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3") +
  labs(title = "Czech Republic: Energy Source Progress (1990-2023)",
       subtitle = "Per-capita energy consumption by source",
       y = "kWh per capita", x = "Year")
```

Interpretation: Czechia shows a strong historical reliance on Coal and Nuclear. While Solar and Wind (at the top) are growing, they still represent a small "wedge" compared to the industrial baseload provided by traditional sources.

10.2. Forecasting Solar Transition via Cluster Logic
Instead of just looking at Czechia's past, we use the "Aggregated Forecast" method. We identify Czechia's cluster and use the cluster's average growth rate to predict Czechia's 2030 solar capacity.

```{r czechia_forecasting, include=TRUE}
# 1. Identify Czechia's current cluster (using the most recent year)
cze_latest <- master_data %>% filter(Entity == "Czechia" & Year == 2024)
target_cluster <- cze_latest$Cluster

# 2. Calculate average growth rate of Solar for the WHOLE cluster (2012-2022)
# This reduces the variance of the forecast (Class 05 logic)
cluster_solar_growth <- master_data %>%
  filter(Cluster == target_cluster & Year >= 2012) %>%
  group_by(Entity) %>%
  summarise(growth_rate = (last(Solar) - first(Solar)) / (last(Year) - first(Year))) %>%
  summarise(avg_cluster_growth = mean(growth_rate, na.rm = TRUE))

# 3. Apply this growth rate to Czechia for the next 6 years (to 2030)
current_solar <- cze_latest$Solar
forecast_years <- 2030 - 2024
projected_solar_2030 <- current_solar + (cluster_solar_growth$avg_cluster_growth *
forecast_years)

# 4. Create the Forecast Plot
forecast_data <- data.frame(
  Year = c(2024, 2030),
  Solar = c(current_solar, projected_solar_2030),
  Label = c("Current (2024)", "Forecast (2030)")
)

ggplot(forecast_data, aes(x = Year, y = Solar)) +
  geom_line(linetype = "dashed", color = "grey", size = 1) +
  geom_point(aes(color = Label), size = 5) +
  theme_minimal() +
  ylim(0, max(projected_solar_2030) * 1.2) +
  labs(title = "Czechia 2030 Solar Forecast (Cluster-Based)",
       subtitle = paste("Based on the trajectory of Cluster", target_cluster),
       y = "Solar kWh per capita")
```

Conclusion: "If the Czech Republic maintains the adoption speed typical of its cluster,

it will reachover 1000 kWh per capita in solar energy by 2030. This provides a data-driven benchmark for national energy goals."