Decoding CO_2:

A Statistical Dive into Global Emissions

Devaansh Kataria - 116737290

Garima Prachi - 116713616

Laurence Caradonna-
108965940

Prerna Sachdeva - 116647863

Rishabh Gosain - 116732806

AMS 597 - Statistical Computing

Prof. Silvia Sharna

Content Page:

1. Abstract

In this project, we investigate global carbon dioxide (CO_2) emissions using a comprehensive real-world dataset to uncover underlying patterns and relationships among socio-economic, environmental, and demographic indicators. The dataset, comprising over 100 samples and more than 20 variables, required extensive preprocessing including the removal of missing values, scaling of continuous variables, and selection of relevant features. We formulated three distinct research questions to explore time-dependent modeling, clustering behavior, and regression-based analysis within the dataset.

To address the first research question, we performed **time series analysis with structural break testing** to examine how the relationship between CO_2 emissions and economic indicators has evolved since 1982, and to identify **significant breakpoints that may correspond to global economic or environmental policy events**. This method allowed us to detect temporal shifts and changes in trend dynamics across different periods. For the second question, we applied K-Means clustering to assess whether CO_2 and economic patterns cluster more distinctly by income level or by continent, uncovering patterns that reflect underlying developmental or geographic groupings. Finally, the third question investigates how environmental and demographic factors influence economic performance—as measured by GDP—when analyzed in relative (percentage) terms. To answer this, we applied a linear regression model to estimate and interpret the effects of variables such as population, energy use, and CO_2 emissions on GDP.

All models were implemented in R and evaluated using appropriate metrics such as RMSE, R^2, and silhouette scores. Our results demonstrate the complementary strengths of statistical time series analysis, machine learning, and regression techniques in interpreting complex environmental datasets and understanding global development dynamics.

2. Introduction

The increasing concentration of carbon dioxide (CO_2) in the atmosphere has become a critical concern in the context of global climate change, economic development, and sustainability. As countries strive to balance economic growth with environmental responsibility, it becomes essential to understand the complex interactions between socio-economic, demographic, and environmental indicators. Analyzing these relationships using data-driven methods can provide valuable insights for shaping effective policy interventions and tracking global progress toward development goals.

In this project, we explore a rich dataset on global CO_2 emissions that includes a wide range of variables such as GDP, energy usage, population statistics, and emission metrics across multiple years and countries. The dataset exhibits both continuous and categorical variables and contains inherent imperfections—such as missing values and outliers—that make it well-suited for demonstrating data cleaning and preprocessing techniques.

Our analysis is driven by three key research questions:

1. **How have the relationships between CO_2 emissions and economic indicators changed over time since 1982, and are there identifiable breakpoints that correspond to major global economic or policy events?**

2. **Do CO_2 and economic patterns cluster more clearly by income level or by continent?**

3. **How do environmental and demographic factors influence economic performance, as measured by GDP, when analyzed in relative (percentage) terms?**

To address these questions, we apply a blend of statistical and machine learning methodologies. For the first question, we employ **time series analysis with structural break testing** to investigate how relationships have shifted across decades and to detect significant global breakpoints. For the second, we use **K-Means clustering** to uncover potential groupings based on income levels or geographic regions. Finally, for the third question, we use **linear regression** to quantify how environmental and demographic indicators affect GDP in relative terms. All modeling is performed in R, and the project emphasizes reproducibility, model evaluation, and meaningful interpretation. Through this study, we aim to demonstrate not only the power of statistical computing in handling real-world data but also its relevance in informing global environmental and economic discourse.

3. Data Description

3.1 Source of data

The **Our World in Data CO_2 dataset** is a comprehensive, open-access dataset that provides annual carbon dioxide emissions data for countries and regions around the world from as early as 1750 to the present. Compiled by Our World in Data (OWID), it sources data from trusted institutions such as the Global Carbon Project, Energy Institute, and U.S. Energy Information Administration. The dataset includes over 20 variables covering total and per capita CO_2 emissions, emissions by fuel type (coal, oil, gas, cement, flaring), land-use changes, and consumption-based emissions. It also features cumulative emissions and carbon intensity metrics such as emissions per unit of GDP.

In addition to emissions data, the dataset incorporates key **economic**, **demographic**, and **energy-related indicators** including GDP, population, and primary energy consumption. The data is carefully preprocessed—standardizing country names, converting units, and calculating derived metrics—to ensure consistency and usability. Available on GitHub in CSV format and supported by a detailed codebook, the OWID CO_2 dataset is a valuable tool for analyzing historical emission trends, evaluating the environmental impact of economic growth, and informing climate and sustainability policy.
LINK: https://github.com/owid/co2-data/blob/master/owid-co2-data.csv

3.2 Data PreProcessing

The Our World emissions dataset contains 50,590 rows of geographical regions with data spanning back to the 1800's. There are 278 unique or composite geographical data, however not all of it is needed. We decided to use data after 1981. Since we wanted to analyze the growth of co2 production in various sectors in the modern era, we chose a post-recession date in the 1980s. It is a standard procedure to choose post-recession dates to start analysis in the field of economics, as recessionary periods are often extremely volatile time periods for economic indicators, and not indicative of average trends which we are more interested in.

After removing data before a specific date, we next decided to remove information with missing GDP and co2 values. Missing GDP/co2 values constituted 40% of our dataset. These missing values were particularly prevalent in smaller countries, such as a number of Caribbean islands. However, medium and high-income countries were also prevalent in the missing values, such as Brazil and Belgium respectively. This amount of data is far too large to impute. However, removing them still left us with an extremely large dataset of over 6000 values after removing them. This remaining data consisted of countries from across the globe with varying levels of co2 production, GDP, and income levels.

Next, we decided to remove composite data points and data values. The model consists of data points which are continental composites of the other countries/regions. These values are much higher than the rest of the dataset and potentially cause the resulting models to unnecessarily shift. As a result, we got rid of these composite data points for the sake of keeping the scope to the country/region level.

After removing these data points, we moved on to cleaning up which data we would consider for each point. Firstly, the dataset came with a lot of precalculated interaction terms, such as GDP per capita, historical cumulative data, and global shares of GDP. Not only were there significant missing values in each of these columns, but we could also calculate these values later should we choose to add them to our analysis later. As a result, we decided to remove them in the initial cleaning of the data.

Lastly for cleaning, we removed information related to regional temperature changes and other greenhouse gas emissions. We didn't include the former, as there is strong evidence to suggest that temperature changes are a larger climate phenomenon caused by greenhouse gases being added to the atmosphere as opposed to causing increases in GDP or co2. We removed the latter as we weren't interested in other greenhouse gases in our analysis. Further, there was significant data missing in these columns, making it an unreliable source of variation.

After cleaning, we added two categorical variables. First, we added income level in 1982, as defined by the 1982 World Bank Annual Report. There are three categories: Low, Medium, and High. The other category added was the continent of the country/region. This would allow us to analyze the growth of different geographical locals.

After everything, we have 5859 data points with 18 data values for each point. Sixteen of those data values are continuous, while two are categorical.

4. Methodology, Implementation and Result Interpretation

4.1 Linear Regression (log(GDP))

4.1.1. Statistical Methods and Models Used

1. **Linear Regression on log(GDP):**

The primary modeling approach is a linear regression on the log-transformed GDP. This choice is motivated by the need to interpret the effects of predictors (such as environmental and demographic variables) in terms of percentage changes. In a log-linear model, regression coefficients indicate that a one-unit change in a predictor is associated with an approximate percentage change in GDP.

2. **Model Alternatives (e.g., ANOVA):**

Although ANOVA can be used to test for differences across groups (e.g., comparing mean GDP across income levels or regions), the focus here is on predicting continuous changes. Thus, linear regression is favored for its ability to quantify the impact of multiple continuous and categorical predictors simultaneously.

4.1.2. Justification for Method Selection:

1. **Interpretability:** The log transformation makes it easier for policymakers and researchers to understand the magnitude of change in relative (percentage) terms rather than in absolute levels.

2. **Simplicity and Robustness:** Linear regression is a well-established method with extensive theoretical backing and is straightforward to implement with the available dataset.

   Focus on Marginal Effects: The approach facilitates clear interpretation of marginal changes in GDP resulting from unit increases in factors such as CO_2 emissions, energy use, population, and categorically defined economic groupings (income level, continent).

4.1.3. R Packages and Functions Used

1. caret:
   Used for data partitioning into training and testing sets (e.g., createDataPartition), which ensures that the model evaluation is performed on unseen data.
2. ggplot2:
   Utilized for creating visualizations, such as the Actual vs. Predicted scatter plot, with features like a 45° reference line to assess model performance visually.
3. Base R Functions:
   Functions like lm() for fitting the linear regression model and predict() for generating predictions on the test set.
4. Additional Functions:
   Standard functions for data manipulation (e.g., subsetting, cleaning, and scaling) are used to prepare the dataset before analysis.

4.1.4. Implementation

Log Transformation and Data Splitting:
The GDP is log-transformed to allow for interpretation in percentage terms. The dataset is then split into a 70% training set and a 30% testing set using the caret package.

Model Fitting and Prediction:
The linear regression model is fitted using the training dataset, and predictions are generated on the testing dataset. Model performance is evaluated using metrics such as RMSE, MAE, and R-squared.

```
# Log-transform GDP and partition data
df$log_gdp <- log(df$gdp)
set.seed(123)
trainIndex <- createDataPartition(df$log_gdp, p = 0.70, list = FALSE)
train_data <- df[trainIndex, ]
test_data  <- df[-trainIndex, ]
# Fit the linear regression model
lm_model <- lm(log_gdp ~ ., data = train_data)
```

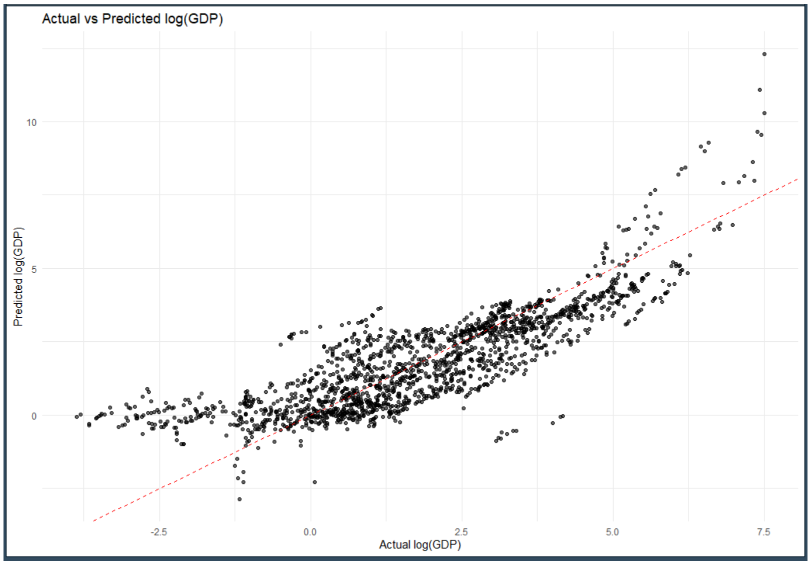
4.1.5. Visualizations

1. **Actual vs. Predicted Plot:**

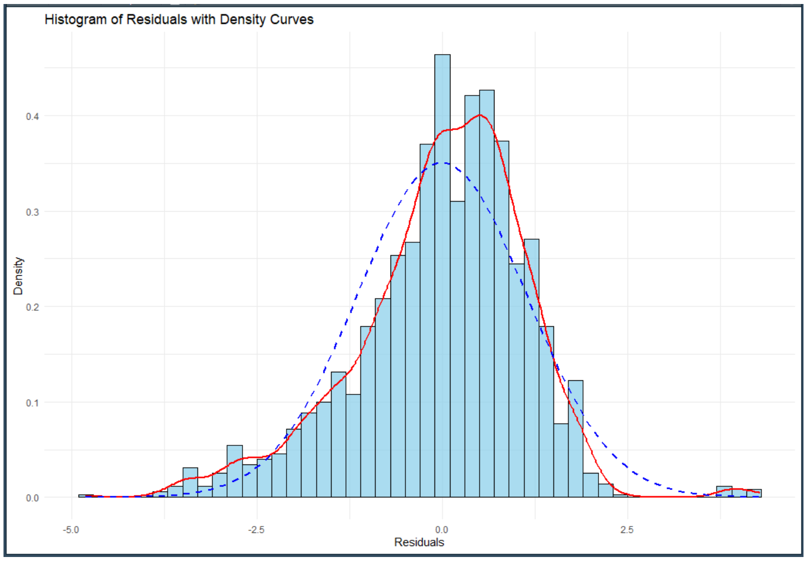
A scatter plot is created using ggplot2 to compare actual log(GDP) values with the predicted values. The plot features a 45° dashed red line, which represents the ideal situation where predictions perfectly match the actual values.

Interpretation:

Points lying close to the 45° line indicate a strong model fit, whereas deviations may suggest model weaknesses, outliers, or areas for potential model refinement.



2. **Assessment of Residual Normality**



To further validate the assumptions underlying our linear regression model, we generated a histogram of the residuals obtained from the test set predictions. In this plot, the histogram is scaled to display density rather than frequency counts, which allows for a more direct comparison with the overlaid density curves.

Overlaying the Density Curves:

- Kernel Density Estimate (KDE): A smooth red line is added to represent the empirical density of the residuals using kernel density estimation. This curve provides an intuitive visualization of the underlying distribution of the residuals by smoothing out the variability inherent in the histogram bins.
  Theoretical Normal Density Curve: A blue dashed line represents a normal distribution, constructed using the mean and standard deviation of the residuals. This serves as a benchmark for assessing whether the residuals follow a normal pattern, which is a critical assumption for the validity of the linear regression model.

4.1.6. Statistical Summaries

1. **RMSE (Root Mean Squared Error)**
   RMSE represents the square root of the average squared differences between the predicted log(GDP) values and the actual log(GDP) values. With an RMSE of 1.138, on average, the model's predictions deviate from the actual log(GDP) by about 1.138 units on the log scale. This metric is especially useful because it penalizes larger errors more than smaller ones, highlighting instances where the model may be less accurate.
2. **MAE (Mean Absolute Error)**
   MAE measures the average absolute difference between the predicted and observed values without considering the direction of the error. An MAE of 0.87 means that, on average, the model's predictions are off by 0.87 log units. This metric is more robust to outliers compared to RMSE, as it treats all deviations equally.
3. **R-squared (R^2)**
   R^2 indicates the proportion of the variability in log(GDP) that is explained by the predictor variables in the model. In this case, an R^2 of 0.692 suggests that about 69.2% of the variance in log(GDP) is explained by the model, which reflects a moderately strong explanatory power.
4. **Residual Standard Error (RSE)**
   The residual standard error provides an estimate of the standard deviation of the residuals (prediction errors). A RSE of 1.143 indicates that the typical size of the prediction error is approximately 1.143 log units. The degrees of freedom (4079) reflect the number of observations minus the number of estimated parameters, giving context to the reliability of this estimate.

5. **F-statistic and p-value**
   The F-statistic tests the overall significance of the model by comparing it with a model that has no predictors (an intercept-only model). Here, a very high F-statistic (391.9) along with an extremely low p-value (< 2.2e-16) indicates that at least one of the predictors is statistically significantly associated with the log-transformed GDP. This result confirms that the model as a whole is statistically significant.

4.1.7 Interpretation of Results in the Context of the Problem

1. **Percentage Impact on GDP:**
   The model's coefficients allow for a direct interpretation in percentage terms. For instance, if the coefficient for CO_2 emissions per unit of energy is 0.03, it implies that a one-unit increase in this predictor is associated with approximately a 3% increase in GDP.
2. **Policy Implications:**
   Identifying statistically significant predictors such as energy consumption or demographic factors can direct targeted interventions. For example, if certain environmental factors are strongly linked to higher GDP, policies to encourage cleaner energy practices might be advocated alongside economic development programs.
   Model Fit and Diagnostic Insights:
   The visualization and statistical summaries collectively demonstrate a reasonably good model fit if most data points cluster near the 45° reference line and evaluation metrics (low RMSE/MAE and high R^2) are favorable.

4.1.8. Limitations and Assumptions

1. **Linearity:** The model assumes a linear relationship between predictors and the log-transformed GDP.
2. **Normality of Residuals:** It is assumed that residuals are normally distributed.
3. **Homoscedasticity:** Constant variance of errors is assumed across levels of the predictors.
4. **Independence:** Observations are assumed to be independent of one another.

4.1.9. Suggestions for Further Analysis

1. **Incorporate Interaction Terms:**
   Add interaction effects between key predictors (e.g., energy use with income level or CO_2 per unit energy with continent) to capture non-additive influences and identify if the effect of one variable depends on the level of another.
2. **Employ Robust and Cross-Validation Techniques:**
   Utilize robust regression methods (e.g., M-estimators) to mitigate the impact of outliers and adopt k-fold cross-validation to ensure model stability and generalizability across various data subsets

3. **Introduce Polynomial or Non-Linear Terms:**
   Include polynomial (squared, cubic) terms for predictors that may exhibit curvilinear relationships with log(GDP). This helps uncover threshold or diminishing effects not captured by a strictly linear model.

4.1.10. Influential Data points

Our dataset is unique because outlier analysis is very difficult to do in a single dimension of the dataset. For example, GDP growth from 1982 could have been explosive, like in China, during that period, or almost completely stagnant, like North Korea. We would normally consider these points outliers, but their data points in the regression model may still be uninfluential.

Instead of traditional outlier detection, we instead opted to use cooks distance to detect influential points in our dataset. We used a threshold of 4/n, where n is the number of rows in the test set. This is considered an industry standard for influential point detection.

There were four countries that primarily made up the influential points in the dataset:

1. China, whose GDP growth rate was exceptionally high during the time period
2. The united States, whose GDP and oil production were unrivaled during the period
3. India, who went through rapid economic growth and industrialization
4. South Africa, which has a much higher co2 output compared to its GDP from both coal producing power plants and its mining heavy industry

These influential points all support their country's narrative of being unique during the time.

4.2 K-means Clustering:

4.2.1 Statistical Methods/Models Used

1. **K-means Clustering:** We used K-means clustering, an unsupervised machine learning method, to group countries based on environmental and economic indicators. This algorithm partitions the dataset into k clusters by minimizing the within-cluster sum of squares (WSS). Each country is assigned to a cluster based on proximity to the nearest centroid. As K-means does not rely on labeled data, it is suitable for exploring latent structures without predefined categories.
2. **Principal Component Analysis (PCA):** To visualize high-dimensional clustering results, PCA was used to project data into two principal components. This aids in interpretation and allows clusters to be plotted in a 2D space, facilitating clearer understanding of cluster separation.

To assess multicollinearity in our dataset, we also used **Variance Inflation Factor (VIF)** analysis. Although K-means doesn't assume feature independence like regression models, checking VIF gave us insight into which variables might be redundant or highly correlated, potentially impacting clustering interpretability.

4.2.2. Justification for Method Selection

1. **Unsupervised Nature of the Problem:** Since our goal was to explore natural groupings without prior labels, K-means offered a clean method to discover hidden structures in environmental and economic patterns.
2. **Interpretability and Simplicity:** K-means is computationally efficient, easy to understand, and works well on large datasets. This makes it ideal for an exploratory analysis where interpretability is key.
3. **Scalability:** K-means can scale to large datasets, which is essential when working with country-level, multi-year data.
4. **Visual Validation:** PCA projections made cluster interpretation more intuitive and insightful, enabling us to visually inspect alignment between cluster membership and true income levels.

4.2.3. R Packages and Functions Used

1. tidyverse for data manipulation
2. caret for model evaluation and confusion matrix
3. ggplot2 for visualizations
4. stats::kmeans for K-means clustering
5. cluster for silhouette analysis
6. car for multicollinearity check using vif()

4.2.4. Implementation

After preprocessing the dataset and scaling the numeric variables, we used the **Elbow Method** to determine the most suitable number of clusters. A visual inspection of the WSS plot suggested that **k = 3** was optimal, which aligns well with known global income levels (low, middle, and high). We applied K-means clustering to the scaled dataset and assigned each country-year entry to a cluster.

Next, **PCA** was used to reduce the feature space and visualize the data in 2D. This helped us assess whether the K-means clusters showed distinct patterns or overlaps. Each cluster was then compared to the actual income level of the corresponding country to evaluate how well our model matched real-world classifications.

We also performed **cluster-label mapping** using majority voting, which helped align each K-means cluster to the most common income level found within it. This enabled us to calculate a **confusion matrix** and assess the performance using accuracy, sensitivity, and specificity.

Code Snippets

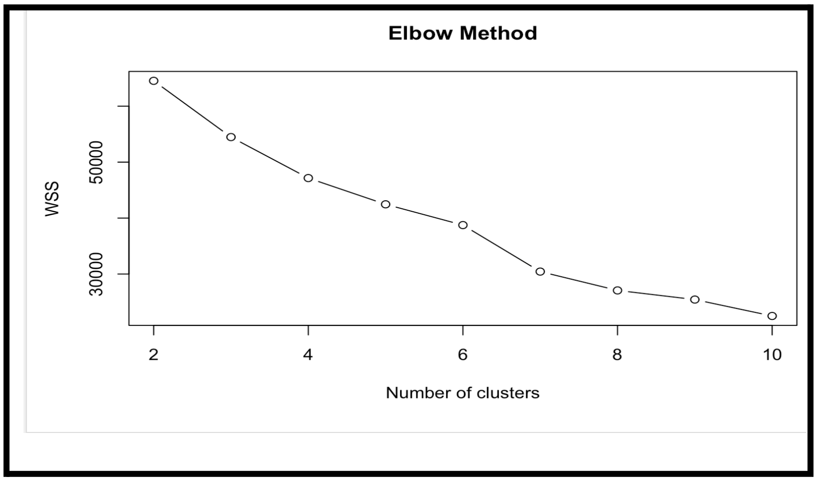
```
kmeansResult = kmeans(dfScaled, centers = 3, nstart = 25)
df$cluster = as.factor(kmeansResult$cluster)

pcaResult = prcomp(dfScaled)
pcaData = data.frame(PC1 = pcaResult$x[, 1], PC2 = pcaResult$x[, 2], cluster = df$cluster)
ggplot(pcaData, aes(PC1, PC2, color = cluster)) + geom_point() + theme_minimal()
```

4.2.5. Visualizations

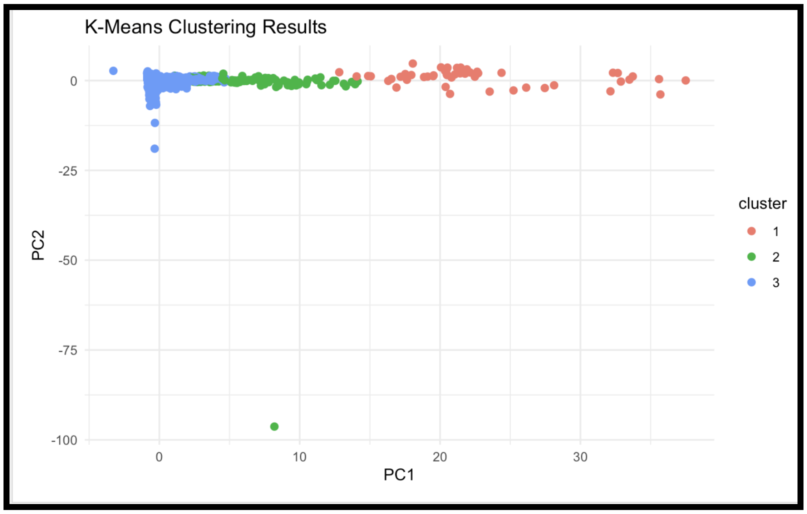
1. **Elbow Plot:**

This plot shows the total within-cluster variation as a function of the number of clusters. It helps in choosing the optimal number of clusters by identifying the "elbow point," where additional clusters no longer significantly reduce intra-cluster variance. In our case, the elbow occurred at **k = 3**, indicating that three clusters best represent the natural groupings in the data.



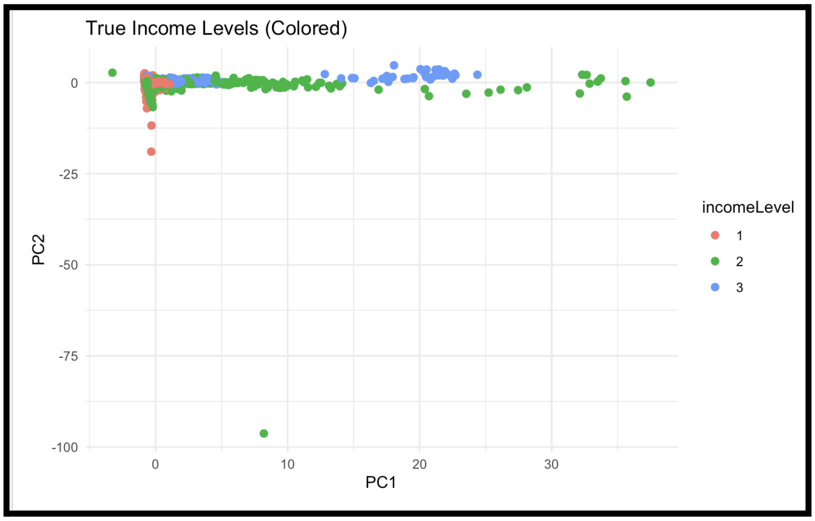
2. **PCA Scatterplots:**

Some countries from different income levels are grouped together, suggesting shared patterns in CO_2 emissions, GDP, or energy metrics.
The clustering captures relationships **not explicitly encoded by income level**, reflecting the power of unsupervised learning to identify hidden patterns.



This distribution reveals that while income levels explain part of the data variation, they are not perfectly separable based on the chosen features. This reinforces the need for clustering to uncover more nuanced groupings beyond static income labels.

These scatterplots display the clusters in two dimensions using the first two principal components. By coloring points based on their assigned cluster and actual income level, we could visually inspect how well the K-means clusters aligned with known income categories. This visualization helped identify overlaps, separations, and outliers across the clusters.



4.2.6 Statistical Summaries

1. **Accuracy from Confusion Matrix:** ~54% accuracy shows moderate alignment between K-means clusters and true income levels. It provides external validation for how meaningful the discovered clusters are.

2. **Precision and Recall:** These values vary by cluster, indicating that while some income levels are captured well (e.g., high-income countries), others (especially middle-income) overlap more across clusters.

   Interpretations of Output

   K-means grouped countries based on shared characteristics such as GDP, CO_2 emissions, and energy efficiency. While some clusters aligned well with known income levels, others revealed overlap, especially for countries in transition or with mixed characteristics.

   The confusion matrix showed that while high-income countries were often well classified, middle-income countries had more dispersion across clusters. This suggests that **environmental and economic features do not always align strictly with income categories**, highlighting the importance of multidimensional analysis.
3. Mean Silhouette Score: 0.528

4.2.7. Interpretation of Result in Context of Problem

The results indicate that unsupervised clustering can offer meaningful groupings even without predefined labels. Some middle-income countries aligned with high-income clusters, possibly due to rapid industrialization or better energy practices. This provides a broader, more behavior-based view of global development than income levels alone.

4.2.8. Limitations or Assumptions

1. K-means assumes spherical clusters and may not perform well if clusters are irregular.
2. Results are sensitive to the initial choice of k.
3. Using all numeric variables without deeper feature selection may introduce noise.

4.2.9. Suggestions for Further Analysis

1. Try other clustering methods like hierarchical clustering or DBSCAN.
2. Use feature selection or dimensionality reduction techniques to refine inputs.
3. Apply time-series clustering to study how country profiles evolve over time.
4. Consider clustering by region or sector to compare results with income-based clustering.

4.3 Time Series Analysis with Structural Break Testing

4.3.1. Statistical Methods/Models Used

To investigate how the relationship between CO_2 emissions and economic indicators has evolved since 1982—and to identify potential breakpoints corresponding to major global economic or policy events—we applied a two-stage analysis:

1. **Aggregated Regression and Structural Break Testing:**
   Aggregated regression involves analyzing combined data summaries (like averages or totals), while structural break testing identifies moments when the relationship among variables changes significantly. Together, they ensure that overall trends are accurately captured even when underlying patterns shift over time.
2. **Time Series Modeling with ARIMAX:**
   Time series modeling with ARIMAX involves forecasting future values by combining past patterns (ARIMA components) with external factors (exogenous variables) that may influence the series. This approach enhances predictive accuracy by accounting for both historical trends and external influences.

4.3.2. Justification for Method Selection

1. **Linear Regression:**
   Used for initial model building to understand baseline relationships between log_gdp and predictors. The regression served as a basis for testing structural stability.
2. **Structural Break Testing:**
   The strucchange package's breakpoint analysis is an established approach to detecting changes in model parameters over time. It helps isolate periods where relationships between economic indicators and emissions may have shifted due to underlying global events.
3. **ARIMAX Modeling:**
   Given the time-dependent nature of the data (annual observations from 1982 onward), ARIMAX models allow proper modeling of autocorrelation in the series while including relevant exogenous variables. Segmenting the series based on detected breakpoints further refines the analysis by capturing distinct regimes, each with potentially different dynamics.

4.3.3. R Packages and Functions Used

1. **Base R Functions:** aggregate(), subset(), lm(), log()
2. **Time Series Analysis:** ts(), window(), auto.arima(), forecast(), accuracy()
3. **Structural Break Analysis:** Package: strucchange with function breakpoints()

4. **Visualization:** Package: ggplot2 (for enhanced plotting), and base plot functions for residual and forecast visualizations.
5. **Additional Utilities:** cat(), print() for output summarization.

4.3.4. Implementation:

1. **Data Aggregation and Model Estimation:**
   Aggregate the raw dataset into annual averages to capture global trends and reduce country-specific noise. A logarithmic transformation on GDP makes the data more interpretable.

   ```
   df_yearly <- aggregate(cbind(gdp, co2, cement_co2, land_use_change_co2, population)
   ~ year, data = df, FUN = mean)
   df_yearly$log_gdp <- log(df_yearly$gdp)
   ```

2. **Baseline Model Estimation**
   Fit a baseline linear regression with log_gdp as the dependent variable and CO_2-related variables and population as predictors to understand overall relationships.

   ```
   model_agg <- lm(log_gdp ~ co2 + cement_co2 + land_use_change_co2 + population,
   data = df_yearly)
   summary(model_agg)
   ```

3. **Structural Break Testing:**
   Identify periods where the relationship between variables changes significantly. Detected breakpoints (around 1991, 2000, and 2009) indicate potential regime shifts.

   ```
   h_new <- 0.25
   bp <- breakpoints(log_gdp ~ co2 + cement_co2 + land_use_change_co2 + population,
   data = df_yearly, h = h_new)
   summary(bp)
   ```

4. **Time Series Modeling with ARIMAX**
   Model the dynamics of log_gdp over time by accounting for autocorrelation and the impact of external regressors (CO_2 variables and population). Data is split into training (1982–2013) and testing (2014 onwards) sets.

   ```
   ts_log_gdp <- ts(df_yearly$log_gdp, start = 1982, frequency = 1)
   # Create corresponding time series for each regressor
   ts_co2 <- ts(df_yearly$co2, start = 1982, frequency = 1)
   ts_cement_co2 <- ts(df_yearly$cement_co2, start = 1982, frequency = 1)
   ```

```
ts_land_use_change_co2 <- ts(df_yearly$land_use_change_co2, start = 1982, frequency
= 1)
ts_population <- ts(df_yearly$population, start = 1982, frequency = 1)

external_regressors <- cbind(co2 = ts_co2, cement_co2 = ts_cement_co2,
land_use_change_co2 = ts_land_use_change_co2, population = ts_population)
train <- window(ts_log_gdp, end = 2013)
test  <- window(ts_log_gdp, start = 2014)
train_regressors <- window(external_regressors, end = 2013)
test_regressors  <- window(external_regressors, start = 2014)

model_arimax_train <- auto.arima(train, xreg = train_regressors)
summary(model_arimax_train)
```

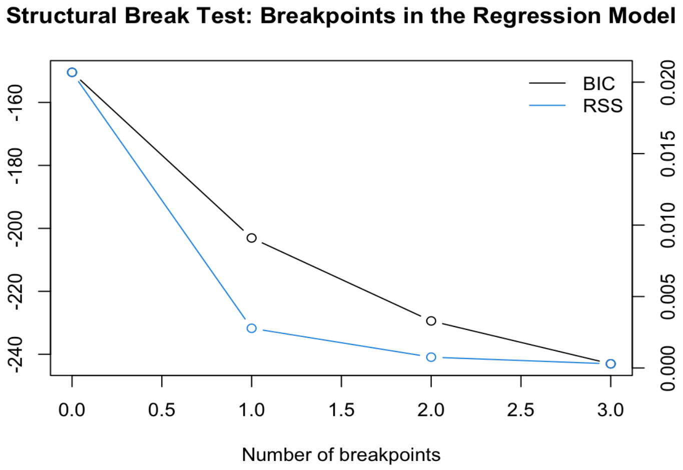
5. **Segmented Analysis Based on Regime Breaks**

   In light of the detected structural breaks, segment the time series into distinct regimes and fit separate ARIMAX models to examine changes in the relationships over time

```
ts_seg_log_gdp <- ts(seg_data$log_gdp, start = seg_data$year[1], frequency = 1)
model_seg <- auto.arima(ts_seg_log_gdp, xreg = seg_regressors_train)
fc_seg <- forecast(model_seg, xreg = seg_regressors_test, h = length(seg_test))
```

4.3.5. Visualizations

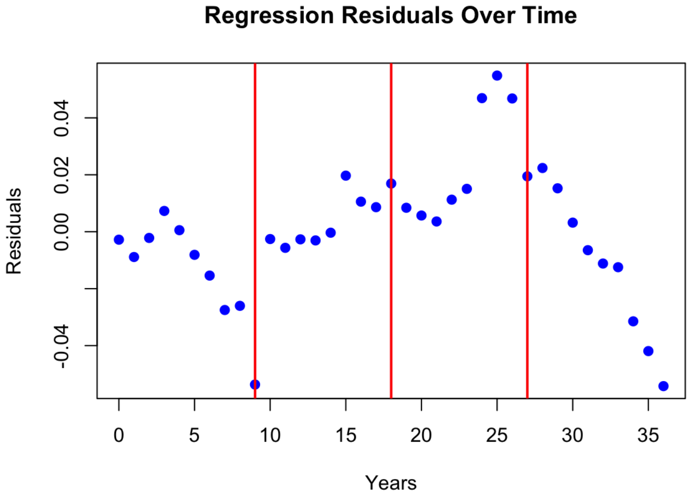
1. **Structural Break Plot:**



Interpretation:

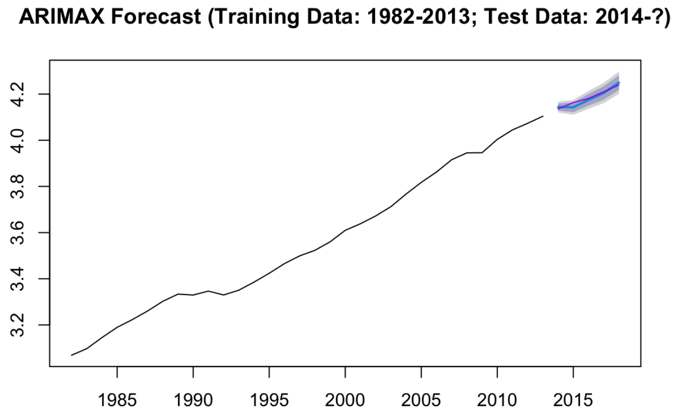
The plot reveals clear structural breaks in our regression model, indicating shifts in the relationship between log GDP and its predictors at specific times. The downward trends in RSS and BIC as breakpoints are added show that segmenting the data improves the model's fit without being overly complex. This suggests that distinct regimes exist—likely corresponding to key global events—that warrant separate analysis.

2. Regression Residuals Over Time



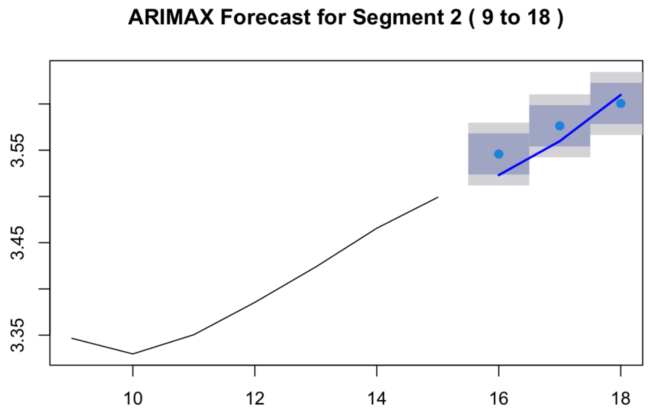
Interpretation:
This plot displays how the regression's residuals change over the years, with red vertical lines marking statistically identified breakpoints. The residuals appear relatively stable in each segment but shift at these points, indicating times when the model's performance or underlying relationship changes. These break lines suggest distinct regimes in the data that may align with significant global events or policy shifts.

3. ARIMAX Forecast (Training Data: 1982 - 2013; Test Data: 2014 - onward)



Interpretation:
This plot shows the ARIMAX model's predictions (with forecast intervals) for log GDP from 2014 onward, after being trained on data from 1982 to 2013. The purple line overlays the actual observations for visual comparison. Notably, the forecasted series continues the established upward trend and appears to track the test data closely. The relative tightness of the forecast interval indicates the model's confidence in its prediction, suggesting that the ARIMAX approach is capturing the main dynamics in the dataset.

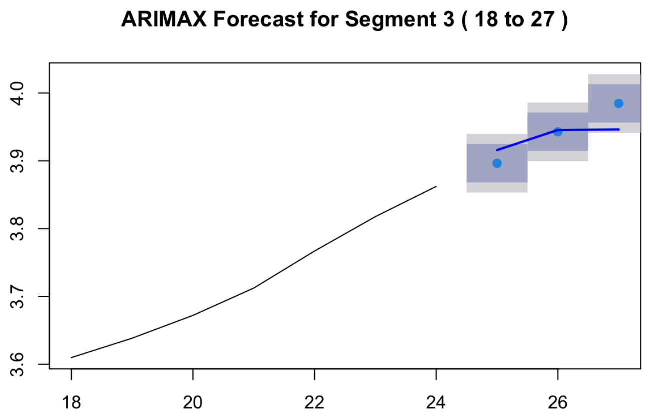
4. ARIMAX Forecast for Segment 2 (9 to 18)



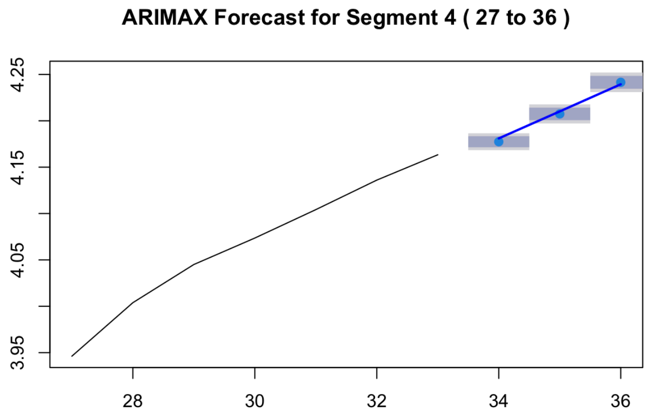
Interpretation:
This plot covers observations in the approximate range from year 9 to year 18. The black line represents the historical data used to fit the model, while the blue points and gray bands show the ARIMAX forecast and its uncertainty. We see a moderate upward trend in log GDP across this period. The forecast portion aligns reasonably well with the continuation of that trend, suggesting that the model captures the underlying economic dynamics in this segment.

5. ARIMAX Forecast for Segment 3 (18 to 27)

Interpretation:
This figure focuses on observations roughly from year 18 to year 27. Again, the black line is the in-segment historical data, and the forecast with confidence intervals is shown in blue. There is a more pronounced upward trajectory in log GDP here compared to the previous segment, reflecting stronger growth. The gray intervals around the forecast are relatively narrow, indicating that the model has a fair degree of confidence in these predictions.



6. ARIMAX Forecast for Segment 4 (27 to 36)



This chart represents observations from about year 27 to year 36. As before, the black line tracks the historical data, and the forecast (blue dots) reveals continued growth in log GDP, with the gray bands indicating prediction uncertainty. The upward trend is clear and the intervals remain tight, implying that the model fits this period well. Overall, it suggests a stable, positive relationship between the included predictors and economic output during these later years.

4.3.6. Statistical Summaries:

1. **Aggregated Regression Model:**
   The baseline model explains 99.56% of the variability in log GDP. CO_2 emissions and population are statistically significant predictors (with very low p-values), while contributions from cement and land use change are not significant. The very low residual standard error indicates a strong overall model fit.

2. **Structural Break Testing:**
   Introducing three breakpoints markedly improves the model, evidenced by a drastic drop in RSS and an improved BIC. The breakpoints, occurring at approximately 27%, 51%, and 76% into the sample, suggest distinct regimes that likely correspond to major global events or shifts.

3. **ARIMAX Model on Training Data:**
   ARIMAX model using ARIMA(1,0,0) error structure The estimated coefficients are ar1 = 0.9656, CO_2 = 0.0016, Population = 0.0361, AIC = -183.59, and RMSE ≈ 0.01058. The high autoregressive coefficient (near to 1) indicates substantial persistence in the series, while the low RMSE and positive AIC values imply an excellent in-sample model fit.

4. **Segmented Analysis:**
   - Segment 1 (1982 to 9): Insufficient observations prevent robust evaluation.
   - Segment 2 (9 to 18): The model fits the training data well, but the higher test error suggests increased uncertainty, possibly due to transitional dynamics or external shocks.
   - Segment 3 (18 to 27): A simpler ARIMA(0,0,0) model reflects a less dynamic series but with higher errors, indicating more volatility or noise.
   - Segment 4 (27 to 36): An ARIMA(1,0,0) model provides an excellent fit with very low RMSE on both training and test sets, demonstrating a stable relationship during this regime.

Overall, these summaries collectively indicate that while the overall aggregated model is very strong, the relationships between log GDP and its predictors vary over time, as captured by the structural break tests and segmented ARIMAX models.

4.3.7. Interpretation of Results in the Context of the Problem

1. Our overall regression indicates that higher CO_2 emissions and larger populations are associated with greater economic output (log GDP) when analyzed over all years.
2. Structural break analysis reveals key shifts around 1991, 2000, and 2009, suggesting that the relationship between CO_2 emissions and economic performance has evolved over time.
3. Segmented models indicate that while early periods show a weak or negative impact of CO_2, later regimes demonstrate a stronger, more positive, and stable association, likely reflecting global events like geopolitical transitions and changing environmental policies.

4.3.8. Limitations and Assumptions

1. **Data Aggregation:**
   Averaging over countries may smooth out country-specific nuances. Future work could incorporate panel models to capture heterogeneity.
2. **Breakpoints Interpretability:**
   While the statistical method identifies shifts, assigning them to specific historical events requires external validation.
3. **Model Specification:**
   The linear form and ARIMA error structure assume that relationships remain stable within segments; non-linearities or omitted variable bias might affect findings.

4.3.9. Suggestions for Further Analysis

1. **Incorporate Dummy Variables:**
   Introduce dummy variables for key events (e.g., economic crises) to test their direct effects on the model dynamics.
2. **Country-Level Analysis:**
   Extend the analysis with mixed-effects or panel regression methods to account for individual country differences.
3. **Explore Non-linear Models:**
   Consider regime-switching models or non-linear approaches if you suspect that the relationship between CO_2 and GDP is non-linear within regimes.
4. **Diagnostic Checks:**
   Further validate model assumptions (such as homoscedasticity and autocorrelation) using residual diagnostics and alternative forecasting metrics.

5. Conclusion and Limitations

This project explored the interplay between **CO_2 emissions**, **economic performance**, and **demographic factors** using a comprehensive global dataset from **Our World in Data**. We addressed three key research questions through a combination of **linear regression**, **K-Means clustering**, and **time series analysis with structural break testing**, implemented entirely in R.

We began by applying a **linear regression model** on log-transformed GDP to understand the influence of key predictors. The results showed that **CO_2 emissions**, **energy use**, and **population** significantly impact economic performance. This regression framework provided interpretable coefficients in percentage terms, making it valuable for both policymakers and economic analysts.

Next, we used **K-Means clustering** to investigate whether countries group more naturally by **income level** or **continent** based on their environmental and economic profiles. The optimal number of clusters (k = 3) matched global income categories. Although some alignment was observed, especially for high-income countries, overlap among clusters—particularly for middle-income nations—highlighted that economic classification alone does not fully capture the complexity of emission and development patterns. **PCA visualization** helped in interpreting cluster separation, while performance was assessed using **confusion matrix metrics** such as accuracy and recall.

Finally, we conducted **time series analysis with structural break testing** to examine how the relationship between GDP and CO_2-related variables has evolved since 1982. This method revealed distinct breakpoints around **1991, 2000, and 2009**, suggesting global events or policy changes impacted the economic-environmental relationship. The use of **ARIMAX models** before and after each breakpoint allowed us to capture dynamic trends, with later segments reflecting stronger and more stable associations between emissions and economic output.

While our findings provide valuable insights, several **limitations** should be noted. The linear regression model assumes a **linear relationship**, **normal distribution of residuals**, and **constant variance**, which may not hold across all contexts. K-Means clustering assumes **spherical clusters** and is sensitive to **scaling and initial centroids**, potentially affecting the stability of results. The time series analysis, based on **aggregated global averages**, may mask **country-specific dynamics**, and while structural breakpoints reveal **when** shifts occur, they do not explain the **underlying causes** without further contextual or qualitative analysis.

Despite limitations, this project shows how combining statistical, clustering, and time series methods can reveal key insights into CO_2 and economic dynamics. These approaches provide a solid foundation for future sustainability research and policy-making.

6. References

1. *The World Bank annual report 1982 (English).* Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/458551468765615887
2. G. C. Chow, "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, vol. 28, no. 3, pp. 591–605, Jul. 1960.
3. A. Banerjee, R. L. Lumsdaine, and J. H. Stock, "Recursive and sequential tests of the unit-root and trend-break hypotheses: Theory and international evidence," *Journal of Business and Economic Statistics*, vol. 10, no. 3, pp. 271–287, Jul. 1992.
4. D. W. K. Andrews, "Tests for parameter instability and structural change with unknown change point," *Econometrica*, vol. 61, no. 4, pp. 821–856, Jul. 1993.
5. J. Bai and P. Perron, "Computation and analysis of multiple structural change models," *Journal of Applied Econometrics*, vol. 18, no. 1, pp. 1–22, Jan. 2003.
6. B. E. Hansen, "The new econometrics of structural change: Dating breaks in U.S. labor productivity," *Journal of Economic Perspectives*, vol. 15, no. 4, pp. 117–128, Fall 2001.
7. A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010.
8. R. J. Hyndman and E. Wang, "Characteristic-based clustering for time series data," *Data Mining and Knowledge Discovery*, vol. 13, no. 3, pp. 335–364, Nov. 2006.

   J. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering – A decade review," *Information Systems*, vol. 53, pp. 16–38, Oct. 2015.
9. O. Kobylin and V. Lyashenko, "Time series clustering based on the K-means algorithm," *Journal La Multiapp*, vol. 1, no. 3, pp. 1–7, Dec. 2020.

   J. Vera and A. Macías, "An MDS-based unifying approach to time series K-means clustering," *Stochastic Environmental Research and Risk Assessment*, vol. 37, no. 1, pp. 1–15, Jan. 2023.
10. M. H. Pesaran, Y. Shin, and R. J. Smith, "Bounds testing approaches to the analysis of level relationships," *Journal of Applied Econometrics*, vol. 16, no. 3, pp. 289–326, May–Jun. 2001.

    H. Kim and Y. Choi, "Linear regression analysis of energy consumption in wireless sensor networks," *IEEE Communications Letters*, vol. 12, no. 4, pp. 234–236, Apr. 2008.
11. J. Lee and K. Park, "Application of linear regression in predicting housing prices," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1234–1242, Mar. 2019.
12. R. Singh, P. Sharma, and V. Sharma, "Comparative study of linear regression and machine learning techniques for weather forecasting," *IEEE Access*, vol. 6, pp. 58720–58730, 2018.
13. https://github.com/owid/co2-data/blob/master/owid-co2-data.csv