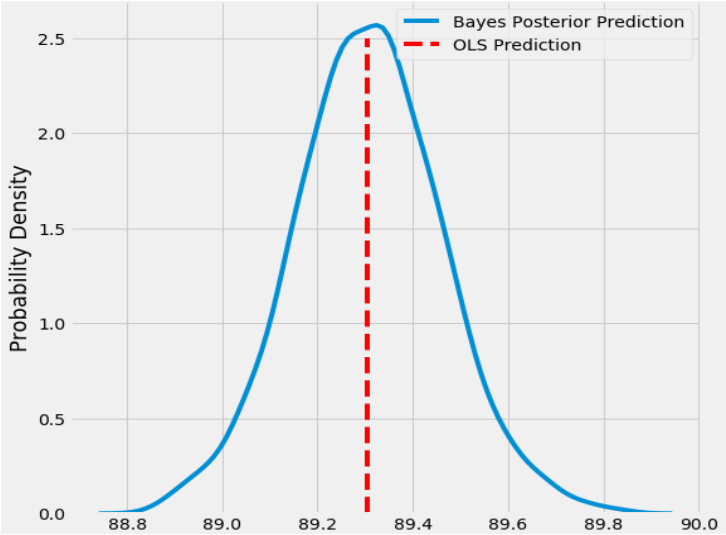

INFO 510 FINAL PROJECT REPORT

Group 4

ABSTRACT

This project will apply Bayesian regression models to understand emission trends and production impacts on global carbon emissions, providing insights into major industries' contribution to climate change.

Arun Koundinya Parasa
Sanjeevteja Ponugumati
Sameer Shoaib

Table of Contents

1. Introduction:

The dataset contains historical records of global greenhouse gas emissions across various commodities, including coal, oil, natural gas, and cement. This project's primary goal is to investigate emissions trends over time and understand the relationship between production levels and emissions for specific commodities. It seeks to provide insights into how industrialization and commodity-specific activities contribute to emissions and how these patterns evolve.

	year	parent_entity	parent_type	commodity	production_value	production_unit	total_emissions_MtCO2e
1	1962	Abu Dhabi National Oil Company	State-owned Entity	Oil & NGL	0.912500	Million bbl/yr	0.3638848
2	1962	Abu Dhabi National Oil Company	State-owned Entity	Natural Gas	1.843250	Bcf/yr	0.1343552
3	1963	Abu Dhabi National Oil Company	State-owned Entity	Oil & NGL	1.825000	Million bbl/yr	0.7277697
4	1963	Abu Dhabi National Oil Company	State-owned Entity	Natural Gas	4.423800	Bcf/yr	0.3224525
5	1964	Abu Dhabi National Oil Company	State-owned Entity	Oil & NGL	7.300000	Million bbl/yr	2.9110786
6	1964	Abu Dhabi National Oil Company	State-owned Entity	Natural Gas	17.326550	Bcf/yr	1.2629390
7	1965	Abu Dhabi National Oil Company	State-owned Entity	Oil & NGL	10.950000	Million bbl/yr	4.3666180
8	1965	Abu Dhabi National Oil Company	State-owned Entity	Natural Gas	25.068200	Bcf/yr	1.8272309
9	1966	Abu Dhabi National Oil Company	State-owned Entity	Oil & NGL	13.505000	Million bbl/yr	5.3854955
10	1966	Abu Dhabi National Oil Company	State-owned Entity	Natural Gas	29.860650	Bcf/yr	2.1765544

The analysis focuses on the following key variables:

- **year:** Observation year.
- **commodity:** Type of commodity (e.g., coal, oil).
- **total_emissions_MtCO2e:** Total emissions in metric tons of CO2 equivalent.
- **production_value:** Production value by commodity.

Two key questions guide this analysis:

1. What are the trends in global emissions over time, and how do different commodities contribute to total emissions?
2. What is the relationship between production levels and emissions for specific commodities, and how does this vary across commodity types?

The findings will help quantify commodity-specific contributions to emissions and identify potential policy targets to reduce greenhouse gas emissions.

2. Question 1 – Global Emission Trends

2.1. Introduction

This question focuses on identifying long-term trends in global greenhouse gas emissions, specifically analysing the contributions of coal, oil, natural gas, and cement. The dataset includes emissions records from industrial and post-industrial periods, enabling us to explore shifts over time.

We are particularly interested in understanding whether emissions growth has plateaued in recent years due to climate policies and which commodities dominate emissions. This analysis will highlight industrialization's impact on emissions patterns and shifts in commodity contributions.

For this part of the analysis, we have used the following variables: **year, commodity** & **total_emissions_MtCO2e.**

2.2. Methods

Model Description: A Bayesian regression model with dummy variables for each commodity type was implemented to evaluate the relationship between emissions (`total_emissions_MtCO2e`) and time (`year`). This model allows us to estimate the contributions of individual commodities to emissions trends.

$$\text{Emissions} \sim \beta 0 + \beta \text{year} \times \text{Year} + \beta \text{commodity} \times \text{Commodity}$$

```
stan_model_code <- "
data {
  int<lower=0> N;               // Number of data points
  int<lower=1> K;               // Number of commodities
  vector[N] year_centered;      // Centered year variable
  matrix[N, K] X_commodity;     // Commodity dummy matrix
  vector[N] emissions;          // Total emissions
}

parameters {
  real beta_0;                  // Intercept
  real beta_year;               // Slope for year
  vector[K] beta_commodity;     // Coefficients for commodities
  real<lower=0> sigma;          // Standard deviation of the error term
}

model {
  beta_0 ~ normal(0, 100);
  beta_year ~ normal(0, 10);
  beta_commodity ~ normal(0, 10);
  sigma ~ cauchy(0, 2);

  emissions ~ normal(beta_0 + beta_year * year_centered + X_commodity * beta_commodity, sigma);
}

generated quantities {
  vector[N] y_rep;
  real log_lik[N];

  for (n in 1:N) {
    y_rep[n] = normal_rng(beta_0 + beta_year * year_centered[n] + dot_product(X_commodity[n], beta_commodity), sigma);
    log_lik[n] = normal_lpdf(emissions[n] | beta_0 + beta_year * year_centered[n] + dot_product(X_commodity[n], beta_commodity), sigma);
  }
}
"
```

Figure 1: Stan Model Code Implementation in R

Priors : Weak priors were chosen to reflect minimal prior knowledge while avoiding overly influential priors:

- Intercept (β0): Normal(0, 100)
- Year effect (βyear): Normal(0, 10)
- Commodity coefficients (βcommodity): Normal(0, 10)
- Error term (σ): Cauchy(0, 2)

MCMC Sampling

- 2000 iterations, 4 chains, and a warm-up of 1000 iterations. Convergence was assessed using trace plots and Rhat diagnostics (Gelman-Rubin statistic).
- Maximum tree depth was set to 10 to ensure efficient exploration of posterior distributions.

Posterior Predictive Checks: Posterior predictive distributions were generated to validate the model by comparing observed and predicted emissions.

Model Validation: Pareto k-diagnostic was used to assess influential data points, while Leave-One-Out Cross-Validation (LOO) provided metrics like elpd_loo.

2.3. Data Analysis and Results.

Model Fitting: The Bayesian regression model was fitted using the `rstan` package, which is already showcased in `Figure 1`.

Diagnostics: Convergence was verified with trace plots and$R$^ values close to 1.

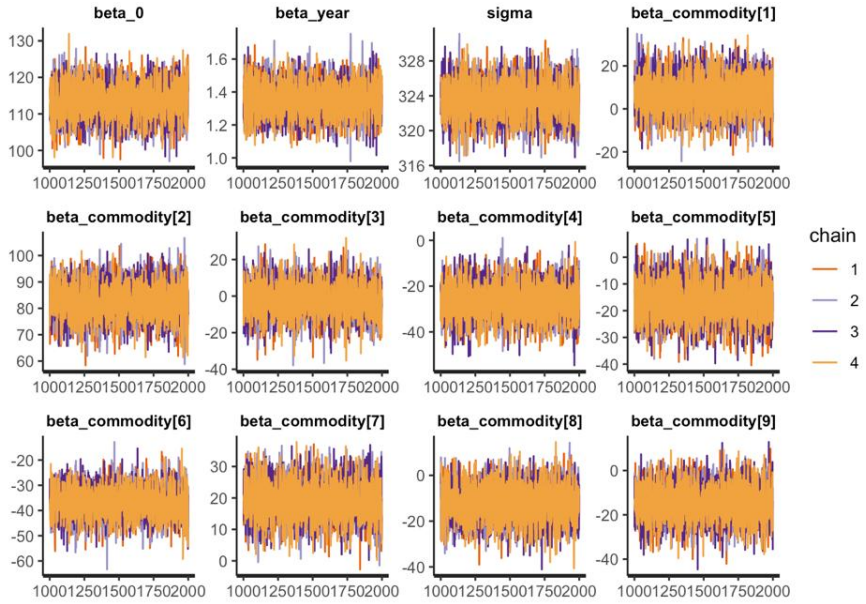

Figure 2: Well-mixed chains and stationarity indicate good convergence

parameters	n_eff	Rhat
beta_0	2,489	1
beta_year	6,266	1
sigma	9,021	1
beta_commodity[1]	5,342	1
beta_commodity[2]	4,239	1
beta_commodity[3]	6,525	1
beta_commodity[4]	5,668	1
beta_commodity[5]	4,999	1
beta_commodity[6]	3,514	1
beta_commodity[7]	3,594	1
beta_commodity[8]	5,928	1
beta_commodity[9]	6,007	1

Table 1: Rhat values are 1

R – code for results snippets.

```
# Model summary
print(fit, pars = c("beta_0", "beta_year", "beta_commodity", "sigma"))

# Posterior predictive checks
ppc_dens_overlay(y = emissions$total_emissions_MtCO2e, yrep =
posterior_predictions[1:100, ])
ppc_stat(y = emissions$total_emissions_MtCO2e, yrep =
posterior_predictions, stat = "mean")

# Convergence diagnostics
traceplot(fit, pars = c("beta_0", "beta_year", "sigma", "beta_commodity"))
```

Model Performance

- *Effective Sample Sizes and Rhat:* The model showed good convergence overall.
- *LOO Cross-Validation:* Leave-One-Out (LOO) cross-validation evaluated the model's predictive performance. The Pareto k-diagnostic values were mainly below 0.7, indicating a good fit.

Results Highlights:

- Significant positive trend for the year effect (βyear), suggesting emissions have increased consistently over time.
- Commodity-specific effects revealed coal as the largest contributor historically, with oil and natural gas rising in recent decades.
- Posterior predictive checks showed good agreement between observed and predicted emissions.

Rmd and html files are attached in the submission folder.

2.4. Discussions

The results indicate a steady rise in global emissions over time, driven by the increasing oil and natural gas contributions, reflecting industrial expansion. While coal remains a dominant contributor, its relative share has declined in recent decades, possibly due to shifts towards cleaner energy sources and climate policies.

This analysis underscores the need for targeted policies to address oil and natural gas emissions while acknowledging coal's historical role. The model also highlights how industrial and economic growth phases influence emissions patterns.

2.5. Future Work

Incorporating Regional Data: Adding geographic dimensions for localized insights.

Improved Priors: Including domain expertise to refine estimates.

Exploring Nonlinear Relationships: To capture potential thresholds or diminishing returns in emissions trends.

3. Question 2 – Production and Emission by Commodity

3.1. Introduction

This question examines how production levels influence emissions across commodity types. We aim to identify whether the relationship between production and emissions varies significantly among commodities like coal, oil, natural gas, and cement.

This analysis will inform commodity-specific strategies to mitigate emissions, particularly for those with higher emissions per unit of production.

For this part of the analysis, we have used the following variables: **production_value, commodity** & **total_emissions_MtCO2e.**

3.2. Methods

Model Description: A Bayesian regression model with dummy variables for each commodity type was implemented to evaluate the relationship between emissions (total_emissions_MtCO2e) and production (production_value).Interaction terms were introduced to capture how the relationship between production and emissions differs across commodities

Emissions∼β0+βprod×Production+βcomm×Commodity Effects+βinter×(Production×Commodity)

```
stan_model_code <- "
data {
  int<lower=0> N;               // Number of data points
  int<lower=1> K;               // Number of commodities
  vector[N] production_value;   // Centered production variable
  matrix[N, K] X_commodity;       // Commodity dummy matrix
  vector[N] emissions;           // Total emissions
}

parameters {
  real beta_0;                 // Intercept
  real beta_prod;               // Slope for production
  vector[K] beta_commodity;       // Coefficients for commodities
  vector[K] beta_interaction;     // Interaction terms for production and commodity
  real<lower=0> sigma;           // Standard deviation of the error term
}

model {
  // Priors
  beta_0 ~ normal(0, 100);
  beta_prod ~ normal(0, 10);
  beta_commodity ~ normal(0, 10);
  beta_interaction ~ normal(0, 10);
  sigma ~ cauchy(0, 2);

  // Likelihood
  emissions ~ normal(beta_0 + beta_prod * production_value +
                X_commodity * beta_commodity +
                production_value .* (X_commodity * beta_interaction), sigma);
}

generated quantities {
  vector[N] y_rep;  // Posterior predictions for emissions
  real log_lik[N];  // Log-likelihood values

  for (n in 1:N) {
    y_rep[n] = normal_rng(beta_0 + beta_prod * production_value[n] +
                    dot_product(X_commodity[n], beta_commodity) +
                    production_value[n] * dot_product(X_commodity[n], beta_interaction), sigma);
    log_lik[n] = normal_lpdf(emissions[n] | beta_0 + beta_prod * production_value[n] +
                      dot_product(X_commodity[n], beta_commodity) +
                      production_value[n] * dot_product(X_commodity[n], beta_interaction), sigma);
  }
}
"
```

Figure 3: Stan Model Code Implementation in R

Priors : Similar priors and sampling parameters as Question 1, with additional terms for interactions:

- Intercept (β0): Normal(0, 100)
- Production Effect (βprod): Normal(0, 10)
- Commodity coefficients (βcommodity): Normal(0, 10)
- Interaction Effect (βinteraction): Normal(0, 10)
- Error term (σ): Cauchy(0, 2)

MCMC Sampling

- 4000 iterations, 4 chains, and a warm-up of 2000 iterations. Convergence was assessed using trace plots and Rhat diagnostics (Gelman-Rubin statistic).
- Maximum tree depth was set to 10 to ensure efficient exploration of posterior distributions.

Posterior Predictive Checks: Posterior predictive distributions were generated to validate the model by comparing observed and predicted emissions.

Model Validation: Pareto k-diagnostic was used to assess influential data points, while Leave-One-Out Cross-Validation (LOO) provided metrics like elpd_loo.

3.3. Data Analysis and Results

Model Fitting: The Bayesian regression model was fitted using the `rstan` package, which is already showcased in `Figure 3`.

Diagnostics: Convergence was verified with trace plots and$R$^ values.

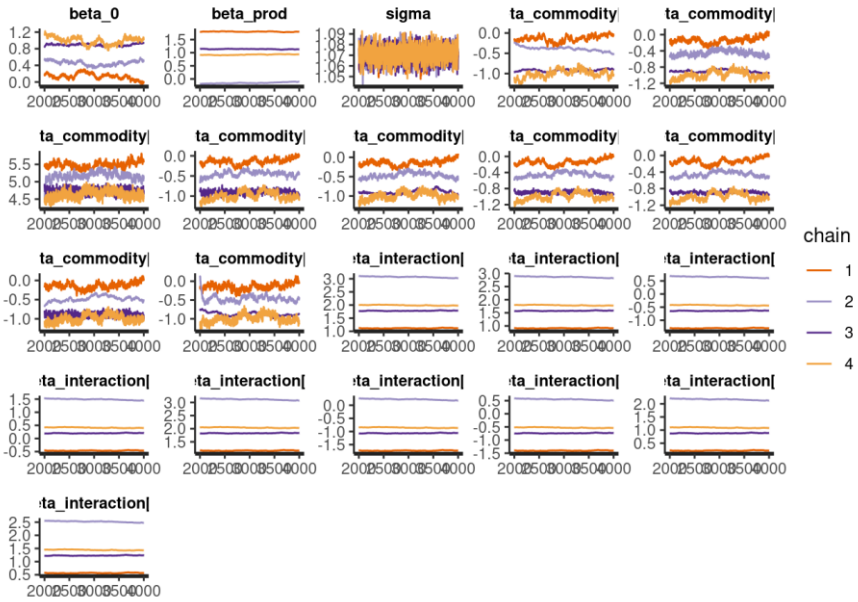

Figure 4: Only a couple of them converge greatly.

parameters	n_eff	Rhat
beta_0	2	6.299313
beta_prod	2	72.450827
sigma	498	1.007015
beta_commodity[1]	2	6.424755
beta_commodity[2]	2	5.83591
beta_commodity[3]	2	4.04803
beta_commodity[4]	2	5.39714
beta_commodity[5]	2	5.480053
beta_commodity[6]	2	5.990681
beta_commodity[7]	2	6.007831
beta_commodity[8]	2	5.225836
beta_commodity[9]	2	4.710203
beta_interaction[1]	2	72.619544
beta_interaction[2]	2	72.445542
beta_interaction[3]	2	72.455616
beta_interaction[4]	2	71.812452
beta_interaction[5]	2	72.391214
beta_interaction[6]	2	72.449801
beta_interaction[7]	2	72.449152
beta_interaction[8]	2	72.303435
beta_interaction[9]	2	72.477782

Table 2: Only Sigma Rhat values are 1

The summary statistics for the effective sample size (n_eff) and R-hat indicated convergence issues for several parameters. Specifically, the interaction terms and production coefficients showed R-hat values significantly higher than the ideal threshold of 1. This suggested that the chains had not mixed properly

R – code for results snippets.

```
# Fit model with interaction term
fit_interaction <- stan(model_code = stan_model_code_with_interaction, data
= stan_data_interaction)

# Diagnostics
traceplot(fit_interaction, pars = c("beta_0", "beta_prod", "beta_comm",
"beta_inter"))

# Posterior predictive checks
ppc_dens_overlay(y = emissions$total_emissions_MtCO2e, yrep =
posterior_predictions[1:100, ])
ppc_stat(y = emissions$total_emissions_MtCO2e, yrep =
posterior_predictions, stat = "mean")
```

Model Performance

- *Effective Sample Sizes and Rhat*: Overall, the model showed reasonable convergence despite some convergence issues. However, the low effective sample sizes and high R-hat values suggest that more iterations are needed for reliable parameter estimates.
- *LOO Cross-Validation*: Leave-One-Out (LOO) cross-validation evaluated the model's predictive performance. The Pareto k-diagnostic values mainly were below 0.7, indicating a good fit, though some values exceeded 0.7, suggesting the need for further model tuning.

Results Highlights:

- The inclusion of interaction terms revealed that the relationship between production levels and emissions varies significantly across commodities. For example, commodities like coal exhibited a stronger positive relationship with emissions, while others like cement showed a more modest effect.
- Posterior predictive checks indicated that the model captured the variability in emissions reasonably well. The density overlay and scatter plots showed that the model's predictions closely matched the observed emissions data.

Note: This MCMC simulation is performed on Cyverse with 64GB RAM and 10 Processors. And anything more than 5000 iterations, Rstudio crashes at this point because of the high number of interaction parameters.

Attachments: Rmd and HTML files showcase the execution output and plots.

3.4. Discussions

The analysis highlighted the significant impact of production levels on emissions, with notable variations across commodity types. The interaction terms helped identify how the effect of production on emissions differs between commodities, offering insights into the role of production strategies in emissions outcomes. This interaction analysis reveals that commodity-specific factors significantly affect emissions intensity.

Given the commodity-specific dynamics, the findings suggest that emissions reduction policies should be tailored to each commodity's unique characteristics. This finding underscores the need for tailored mitigation strategies for high-emission commodities like coal and oil. Cement and natural gas, while lower emitters per unit of production, still require attention to achieve overall emissions reduction targets.

3.5. Future Work

Model Refinement: Future work should involve refining the model to address convergence issues, possibly by adjusting priors or using a more robust method for handling interaction terms.

Policy Implications: Further analysis can be conducted to explore policy levers for reducing emissions, focusing on the specific dynamics identified in this study.

4. Conclusion

- **Global Trends and Commodity Contributions**: The analysis of global emissions trends highlights a consistent increase over time, with coal historically dominating emissions. However, oil and natural gas contributions have grown significantly in recent decades, likely due to industrial expansion and increased energy demands. While smaller in scale, cement emissions are rising in developing regions as infrastructure projects expand. These findings emphasize the dynamic nature of commodity contributions to global emissions and the shifting energy landscape driven by economic and industrial growth.
- **Production-Emission Relationship**: As the interaction analysis demonstrates, the relationship between production levels and emissions varies significantly across commodities. Commodities like coal and oil exhibit higher emissions per unit of production than natural gas and cement, highlighting their outsized contribution to greenhouse gas emissions. This reinforces the importance of tailoring mitigation strategies to specific commodity characteristics rather than adopting uniform approaches.
- **Logical Link Between Questions**: The insights from both questions provide a cohesive narrative. While global trends show the evolving contributions of commodities to emissions, the production-emission relationship analysis delves deeper into explaining the mechanisms behind these trends. The interaction effects indicate that the production scale alone does not fully account for emissions; commodity-specific factors play a crucial role. For example, despite the recent plateau in global emissions growth, rising oil and natural gas production continues to offset potential reductions from cleaner technologies.
- **Policy Implications**: Addressing global emissions effectively requires a macro-level understanding of historical and current trends (Question 1) and a micro-level focus on production-emission relationships (Question 2). Policies should target high-emission commodities like coal and oil with stricter regulations and incentives for cleaner alternatives while leveraging advancements in lower-emission commodities like natural gas and cement to achieve overall reductions.