Final Project

Rahmat

2025-12-05

```r
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
```

Introduction

Climate change and global warming are driven largely by anthropogenic emissions of carbon dioxide (CO_2). Understanding how national energy use and economic activity translate into CO_2 emissions is essential for designing effective climate and environmental policies. In this project, I analyze country-level CO_2 emissions per capita and their relationship with key economic and energy-related predictors.

The data were obtained from the “CO_2 and Greenhouse Gas Emissions’ ’ collection maintained by Our World in Data. The original dataset contains 25,204 observations and 58 variables and includes annual information on CO_2 emissions, greenhouse gases, energy use, and socio-economic indicators for many countries and territories. The main goal of this project is to determine whether CO_2 emissions per capita can be effectively modeled and predicted using a small set of interpretable economic and energy-use variables.

I chose this topic because it aligns closely with my long-term interest in environmental health, pollution, and the role of data-driven methods in informing public policy. As a statistics student interested in environmental applications, this project provides an opportunity to connect statistical modeling with a substantive problem that I care about: how energy consumption and economic growth contribute to global emissions.

To demonstrate that the dataset is worth analyzing, I first explore its structure, variable types, and degree of missingness. I then focus on a subset of variables that are directly related to CO_2 emissions per capita, energy use, and economic scale. The remainder of the report presents a detailed data evaluation, modeling strategy, analysis of results, and a discussion of the strengths and limitations of the chosen methods.

Data Evaluation and Preparation

Original Dataset and Cleaning Steps

The original dataset contains 25,204 rows and 58 columns, covering multiple countries and years. However, many variables for earlier years suffer from substantial missingness, particularly for components of greenhouse gas emissions and energy-use indicators. To obtain a cleaner subset for modeling, I applied the following steps:

- Restricted the analysis to years ≥ 1990, where CO_2 and energy-related variables are more consistently reported.
- Selected 11 variables related to emissions, economic activity, and energy use.
- Removed rows with missing values in any of the selected variables.

After these steps, the cleaned dataset used for analysis consists of 2,299 observations and 11 variables. Table~1 summarizes the change in dataset size.

Dataset	Rows	Columns
Original (all years, all variables)	25,204	58
Cleaned (years $\geq$ 1990, selected vars, complete cases)	2,299	11

Table 1: Summary of dataset size before and after cleaning.

Variables Used in the Analysis

The final modeling dataset includes one country identifier, one time variable (year), and nine continuous predictors related to CO_2 emissions, population, GDP, and energy use. Table~2 lists each variable, its description, and type.

Variable	Description	Type
`country`	Country or territory name	Categorical
`year`	Calendar year	Numeric
`co2_per_capita`	CO_2 emissions per capita (tonnes)	Numeric
`gdp`	Gross domestic product (constant dollars)	Numeric
`population`	Total population	Numeric
`coal_co2`	CO_2 emissions from coal	Numeric
`oil_co2`	CO_2 emissions from oil	Numeric
`gas_co2`	CO_2 emissions from natural gas	Numeric
`primary_energy_consumption`	Total primary energy consumption	Numeric
`energy_per_capita`	Energy use per person	Numeric
`energy_per_gdp`	Energy use per unit of GDP	Numeric

Table 2: Variables used in the modeling dataset, with descriptions and types.

Exploratory Analysis and Transformations

Exploratory plots indicated that several key variables, including CO_2 emissions per capita, GDP, population, and the fuel-specific emissions (coal, oil, gas), were strongly right-skewed. In addition, a correlation heatmap showed substantial correlation among the energy and emissions variables, suggesting multicollinearity.

To address skewness and make relationships more linear, I applied log-transformations to CO_2 emissions per capita, GDP, population, coal, oil, gas, total energy consumption, energy per capita, and energy per GDP. The response variable for modeling is therefore

$$\texttt{log\_co2\_pc} = \log(\texttt{co2\_per\_capita} + 1),$$

with similar log-transformations defined for the predictors (a small constant is added where needed to avoid taking the log of zero).

The combination of data cleaning, restriction to modern years, and log-transformations yields a relatively large, clean, and well-behaved dataset suitable for regression and machine learning models.

Modeling Approach

Train–Test Split

To evaluate predictive performance, the cleaned and transformed dataset was randomly split into an 80% training set and a 20% test set. All model fitting and tuning were carried out on the training data only. The response variable for both models was the log-transformed CO_2 emissions per capita (`log_co2_pc`), and the predictors were the corresponding log-transformed versions of GDP, population, coal CO_2, oil CO_2, gas CO_2, total energy consumption, energy per capita, and energy per GDP.

Modeling Framework

Two modeling approaches were selected:

- **Ridge Regression (Interpretable Model):** Ridge regression incorporates an ℓ_2 penalty that shrinks coefficients toward zero without eliminating predictors. This penalty stabilizes coefficient estimates in the presence of multicollinearity and allows all predictors to remain in the model, which is desirable when each variable represents a meaningful aspect of national energy use or emissions. The ridge penalty parameter λ was selected using 10-fold cross-validation on the training set.
- **Random Forest (Predictive Model):** Random forests are ensemble tree models that capture nonlinear relationships and interactions between predictors. They are relatively robust to skewness and multicollinearity and often provide strong predictive performance. A 10-fold cross-validated random forest was fit using 500 trees, a node size of 5, and `mtry` = 3 predictors considered at each split. Test-set performance was used to assess out-of-sample accuracy.

The next sections present the fitted models, cross-validation results, and comparative performance of the ridge and random forest models.

Ridge Regression Details

The ridge model was implemented using the `glmnet` package. The predictors were assembled into a model matrix, and `glmnet` automatically standardized each predictor (subtracting its mean and dividing by its standard deviation) before fitting the penalized regression. A sequence of candidate λ values was considered, and 10-fold cross-validation was used to estimate the mean squared error for each value of λ.

The optimal penalty parameter, denoted λ_{min}, was chosen as the value that minimized the cross-validated error. The final ridge model was then refit on the full training data using this value of λ_{min}, and predictions were generated for both the training and test sets.

Random Forest Details

The random forest model was implemented using the `randomForest` and `caret` packages. A 10-fold cross-validation procedure was used to tune the hyperparameter `mtry`, which controls the number of predictors randomly sampled at each split in each tree. The final model used `mtry` = 3, 500 trees (`ntree` = 500), and a minimum node size of 5. These settings balance predictive accuracy and computational cost.

Random forest predictions for the test set were obtained by averaging predictions across all trees. In addition, a variable importance plot was produced to identify which predictors contributed most strongly to reducing prediction error.

Analysis of Results

This section evaluates the predictive performance of the ridge regression and random forest models using the independent test set. Two metrics were used for assessment: the Root Mean Squared Error (RMSE), which penalizes larger errors more heavily, and the Mean Absolute Error (MAE), which measures the average magnitude of prediction error. Lower values of both metrics indicate better predictive accuracy.

Test Set Performance

Table~3 summarizes the test-set RMSE and MAE for both models. These values provide a direct comparison of how well each approach generalizes to new data.

Model Comparison

The results reveal substantial differences between the two models. Ridge regression, a linear shrinkage method, achieved a test RMSE of 0.22496 and a test MAE of 0.16849. The closeness of its training and test errors

Model	RMSE (Test)	MAE (Test)
Ridge Regression	0.22496	0.16849
Random Forest	0.06546	0.03953

Table 3: Comparison of test-set predictive accuracy for ridge regression and random forest.

indicates that the ridge model does not overfit and provides a stable linear approximation of the relationship between log CO_2 emissions and the predictors.

Random forest, however, demonstrated considerably stronger predictive performance, with a test RMSE of 0.06546 and a test MAE of 0.03953. These values are markedly lower than those of ridge regression, suggesting that the relationships between national CO_2 emissions and the energy and economic predictors exhibit meaningful nonlinearities and interactions that a linear model cannot fully capture.

Interpretation of Error Metrics

RMSE penalizes large errors more severely than MAE. The random forest model achieves both lower RMSE and lower MAE, indicating that it not only reduces large prediction mistakes but also improves overall accuracy across all observations. Ridge regression performs reasonably well but displays higher systematic error, suggesting that its linear structure is too restrictive to fully capture the variation in CO_2 emissions per capita.

Variable Importance in the Random Forest

The variable importance plot for the random forest model shows which predictors contribute most strongly to reducing prediction error. Energy use per capita and fossil fuel emissions (especially coal and oil) emerge as highly important predictors, reflecting their central role in determining CO_2 emissions per capita. Population and GDP per capita also contribute, but their importance is lower once energy and fuel-specific emissions are included.

Discussion of Final Models and Analysis

This section presents the final ridge regression model, interprets its coefficients, and summarizes the substantive and predictive conclusions drawn from both modeling approaches.

Final Ridge Regression Model

Using 10-fold cross-validation, the optimal penalty parameter for ridge regression was identified as $\lambda_{\min}$. The corresponding coefficient estimates indicate that the model relies primarily on the log-transformed predictors. The final ridge model can be written as:

$$\begin{aligned}\widehat{\text{log\_co2\_pc}} &= \beta_0 + \beta_1 \cdot \text{log\_gdp} + \beta_2 \cdot \text{log\_pop} + \beta_3 \cdot \text{log\_coal} \\ &\quad + \beta_4 \cdot \text{log\_oil} + \beta_5 \cdot \text{log\_gas} + \beta_6 \cdot \text{log\_energy} \\ &\quad + \beta_7 \cdot \text{log\_epc} + \beta_8 \cdot \text{log\_epg}.\end{aligned} \tag{1}$$

Because the model is log–log, each coefficient represents an elasticity: the percent change in CO_2 emissions per capita associated with a 1% change in the predictor, holding all other predictors constant.

- **Energy per Capita ($\beta = 0.314$):** A 1% increase in energy consumption per person is associated with an estimated 0.31% increase in CO_2 emissions per capita. This is the strongest effect and underscores the central role of national energy intensity in driving emissions.

- **Coal, Oil, and Gas CO_2 ($\beta = 0.106, 0.092, 0.066$):** Each fuel source positively contributes to emissions. A 1% increase in coal-related emissions increases total emissions by about 0.11%, oil by 0.09%, and gas by 0.07%. These results reflect the high carbon intensity of fossil fuel consumption.
- **Population ($\beta = -0.152$):** Larger national populations are associated with lower CO_2 emissions per capita. This reflects global patterns where densely populated countries tend to have lower per-person emissions than small, wealthy nations.
- **GDP ($\beta = -0.031$):** After controlling for energy use and fossil fuel emissions, higher GDP per capita is slightly associated with lower emissions, suggesting that economic development may accompany modest improvements in efficiency.
- **Energy per GDP ($\beta \approx -0.001$):** More energy-efficient economies exhibit slightly lower emissions, though the effect is small once other predictors are included.

Summary of Model Performance

Ridge regression offers interpretability, stable coefficient estimates, and insight into the relative influence of each predictor. However, its linear structure limits its predictive power when relationships are nonlinear. Random forest provides substantially higher predictive accuracy because it models complex patterns and interactions automatically. The trade-off is interpretability: although variable importance plots identify influential predictors, the model does not yield an explicit equation.

If the goal is prediction, random forest is the superior model. If the goal is interpretation and understanding, ridge regression provides meaningful elasticity-based insights that connect directly to policy-relevant quantities.

Conclusion

This project examined the economic and energy-related drivers of national CO_2 emissions per capita using a large international dataset. Through exploratory data analysis, transformations, and two complementary modeling approaches, I gained a deeper understanding of both the structure of the data and the capabilities of different statistical learning methods.

From a substantive perspective, the analysis highlights the central role of energy use per capita and fossil fuel consumption in driving emissions. Coal, oil, and gas all contribute positively to CO_2 emissions per capita, while measures such as population and GDP per capita suggest that some countries achieve modest efficiency gains as they grow wealthier or more densely populated. These findings are consistent with the idea that energy intensity and fuel mix are key levers for reducing emissions.

From a modeling perspective, the project illustrated the strengths and weaknesses of ridge regression and random forest. Ridge regression provided an interpretable linear model that handled multicollinearity and yielded elasticity-based interpretations for each predictor. Random forest, in contrast, delivered substantially lower prediction errors by capturing nonlinearities and interactions that the linear model could not represent.

One important limitation encountered in this project was the presence of missing data, particularly in earlier years and for some energy-related variables. Restricting the analysis to years ≥ 1990 and using complete cases improved data quality but reduced the sample size and limited the historical scope of the analysis. In future work, it would be valuable to explore principled methods for handling missing data (such as multiple imputation) or to incorporate additional data sources to obtain more complete time series.

If given more time, several extensions would be worthwhile: exploring additional nonlinear or semiparametric models (such as gradient boosting or generalized additive models), incorporating temporal or country-specific effects through mixed models or panel-data methods, and examining how policy interventions or technological changes affect emissions over time.

Overall, the project provided meaningful insights into the drivers of CO_2 emissions while illustrating the contrasting strengths of interpretable statistical models and modern machine learning methods. The combination of careful data preparation, appropriate transformations, and complementary modeling tools resulted in a richer understanding of both the data and the global forces shaping carbon emissions.

Appendix

```
## [1] 25204    58
```

```
## Rows: 25,204
## Columns: 58
## $ iso_code                            <chr> "AFG", "AFG", "AFG", "AFG", "AFG",~
## $ country                             <chr> "Afghanistan", "Afghanistan", "Afg~
## $ year                                <dbl> 1949, 1950, 1951, 1952, 1953, 1954~
## $ co2                                 <dbl> 0.015, 0.084, 0.092, 0.092, 0.106,~
## $ consumption_co2                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ co2_growth_prct                     <dbl> NA, 475.00, 8.70, 0.00, 16.00, 0.0~
## $ co2_growth_abs                      <dbl> NA, 0.070, 0.007, 0.000, 0.015, 0.~
## $ trade_co2                           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ co2_per_capita                      <dbl> 0.002, 0.011, 0.012, 0.012, 0.013,~
## $ consumption_co2_per_capita          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ share_global_co2                    <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00~
## $ cumulative_co2                      <dbl> 0.015, 0.099, 0.191, 0.282, 0.388,~
## $ share_global_cumulative_co2         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ co2_per_gdp                         <dbl> NA, 0.009, 0.010, 0.009, 0.010, 0.~
## $ consumption_co2_per_gdp             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ co2_per_unit_energy                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ coal_co2                            <dbl> 0.015, 0.021, 0.026, 0.032, 0.038,~
## $ cement_co2                          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ flaring_co2                         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ gas_co2                             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ oil_co2                             <dbl> NA, 0.063, 0.066, 0.060, 0.068, 0.~
## $ other_industry_co2                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ cement_co2_per_capita               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ coal_co2_per_capita                 <dbl> 0.002, 0.003, 0.003, 0.004, 0.005,~
## $ flaring_co2_per_capita              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ gas_co2_per_capita                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ oil_co2_per_capita                  <dbl> NA, 0.008, 0.008, 0.008, 0.008, 0.~
## $ other_co2_per_capita                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ trade_co2_share                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ share_global_cement_co2             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ share_global_coal_co2               <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00~
## $ share_global_flaring_co2            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ share_global_gas_co2                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ share_global_oil_co2                <dbl> NA, 0.00, 0.00, 0.00, 0.00, 0.00, ~
## $ share_global_other_co2              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ cumulative_cement_co2               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ cumulative_coal_co2                 <dbl> 0.015, 0.036, 0.061, 0.093, 0.131,~
## $ cumulative_flaring_co2              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ cumulative_gas_co2                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ cumulative_oil_co2                  <dbl> NA, 0.063, 0.129, 0.189, 0.257, 0.~
## $ cumulative_other_co2                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ share_global_cumulative_cement_co2  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ share_global_cumulative_coal_co2    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ share_global_cumulative_flaring_co2 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ share_global_cumulative_gas_co2     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ share_global_cumulative_oil_co2     <dbl> NA, 0.00, 0.00, 0.00, 0.00, 0.00, ~
## $ share_global_cumulative_other_co2   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ total_ghg                           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ ghg_per_capita                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ methane                             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ methane_per_capita                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ nitrous_oxide                       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ nitrous_oxide_per_capita            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ population                          <dbl> 7624058, 7752117, 7840151, 7935996~
## $ gdp                                 <dbl> NA, 9421400000, 9692280000, 100173~
## $ primary_energy_consumption          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ energy_per_capita                   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ energy_per_gdp                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
```

Table 4: Data summary

Name	co2_raw
Number of rows	25204
Number of columns	58
Column type frequency:	
character	2
numeric	56
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
iso_code	3256	0.87	3	8	0	219	0
country	0	1.00	4	32	0	244	0

Variable type: numeric

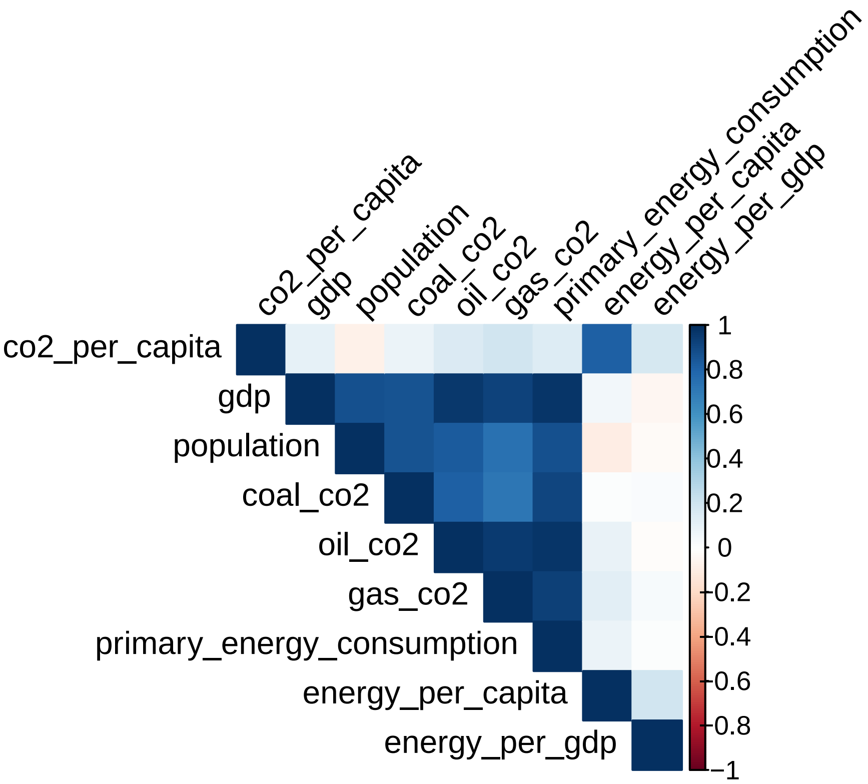
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
year	0	1.00	1.953230e+03	5.379000e+01	1750.00	1925.00	1.967000e+03	1.995000e+03	2.020000e+03
co2	1255	0.95	2.678600e+02	1.521680e+03	0.00	0.53	4.860000e+00	4.282000e+01	3.670250e+04
consumption_co2	21228	0.16	9.167600e+02	3.273350e+03	0.20	10.32	5.709000e+01	2.763800e+02	3.670250e+04
co2_growth_prct	273	0.99	2.110000e+01	7.125700e+02	-99.64	-0.45	3.350000e+00	1.046000e+01	1.023185e+05
co2_growth_abs	1619	0.94	5.150000e+00	5.026000e+01	-1895.24	-0.01	6.000000e-02	1.100000e+00	1.736260e+03
trade_co2	21228	0.16	-2.420000e+00	1.824400e+02	-1658.00	-0.89	1.950000e+00	9.700000e+00	1.028490e+03
co2_per_capita	1897	0.92	4.170000e+00	1.491000e+01	0.00	0.25	1.250000e+00	4.660000e+00	7.486400e+02
consumption_co2_per_capita	21228	0.16	6.570000e+00	6.930000e+00	0.06	1.24	4.360000e+00	9.850000e+00	5.779000e+01
share_global_co2	1255	0.95	4.980000e+00	1.770000e+01	0.00	0.01	6.000000e-02	6.000000e-01	1.000000e+02
cumulative_co2	1255	0.95	1.035710e+04	6.420603e+04	0.00	6.99	9.132000e+01	1.147510e+03	1.696524e+06
share_global_cumulative_co2	1255	0.95	5.130000e+00	1.848000e+01	0.00	0.00	3.000000e-02	4.100000e-01	1.000000e+02

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
co2_per_gdp	9815	0.61	4.200000e-01	4.800000e-01	0.00	0.14	2.800000e-01	5.300000e-01	7.780000e+00
consumption_co2_per_gdp	21443	0.15	3.700000e-01	2.700000e-01	0.01	0.22	3.200000e-01	4.500000e-01	3.540000e+00
co2_per_unit_energy	16063	0.36	2.400000e-01	2.300000e-01	0.00	0.18	2.200000e-01	2.600000e-01	4.640000e+00
coal_co2	8016	0.68	1.753600e+02	7.861100e+02	0.00	0.32	3.980000e+00	3.553000e+01	1.506290e+04
cement_co2	12956	0.49	1.289000e+01	7.726000e+01	0.00	0.13	5.600000e-01	2.900000e+00	1.626370e+03
flaring_co2	20822	0.17	1.500000e+01	4.147000e+01	0.00	0.25	2.070000e+00	1.260000e+01	4.350300e+02
gas_co2	16359	0.35	1.087500e+02	4.410600e+02	0.00	0.38	4.200000e+00	3.083000e+01	7.553390e+03
oil_co2	4665	0.81	1.062500e+02	6.226800e+02	0.00	0.31	2.100000e+00	1.737000e+01	1.222964e+04
other_industry_co2	23205	0.08	1.575000e+01	3.939000e+01	0.00	0.75	2.860000e+00	9.900000e+00	3.038600e+02
cement_co2_per_capita	12986	0.48	1.100000e-01	1.500000e-01	0.00	0.02	7.000000e-02	1.600000e-01	2.740000e+00
coal_co2_per_capita	8344	0.67	1.550000e+00	2.550000e+00	0.00	0.05	4.400000e-01	2.150000e+00	3.418000e+01
flaring_co2_per_capita	20823	0.17	8.800000e-01	5.480000e+00	0.00	0.02	7.000000e-02	2.000000e-01	9.471000e+01
gas_co2_per_capita	16369	0.35	1.410000e+00	3.650000e+00	0.00	0.03	2.800000e-01	1.440000e+00	5.248000e+01
oil_co2_per_capita	5023	0.80	2.640000e+00	1.513000e+01	0.00	0.12	6.300000e-01	2.470000e+00	7.486400e+02
other_co2_per_capita	23205	0.08	8.000000e-02	6.000000e-02	0.00	0.04	7.000000e-02	1.100000e-01	3.600000e-01
trade_co2_share	21228	0.16	2.296000e+01	4.506000e+01	-96.76	-1.76	1.168000e+01	3.638000e+01	3.661500e+02
share_global_cement_co2	12956	0.49	4.420000e+00	1.595000e+01	0.00	0.05	2.000000e-01	1.000000e+00	1.000000e+02
share_global_coal_co2	8016	0.68	6.990000e+00	2.076000e+01	0.00	0.01	1.100000e-01	1.250000e+00	1.000000e+02
share_global_flaring_co2	20822	0.17	5.860000e+00	1.487000e+01	0.00	0.09	7.000000e-01	4.440000e+00	1.000000e+02
share_global_gas_co2	16359	0.35	5.410000e+00	1.841000e+01	0.00	0.03	2.000000e-01	1.250000e+00	1.000000e+02
share_global_oil_co2	4665	0.81	2.990000e+00	1.202000e+01	0.00	0.01	8.000000e-02	5.500000e-01	1.000000e+02
share_global_other_co2	23205	0.08	1.430000e+01	2.857000e+01	0.00	0.30	1.340000e+00	1.009000e+01	1.000000e+02
cumulative_cement_co2	12956	0.49	3.077600e+02	1.706160e+03	0.00	1.61	1.045000e+01	6.646000e+01	4.316319e+04
cumulative_coal_co2	8016	0.68	8.791770e+03	3.913172e+04	0.00	5.52	9.818000e+01	1.248780e+03	7.833620e+05
cumulative_flaring_co2	20822	0.17	4.257000e+02	1.320910e+03	0.00	4.07	4.561000e+01	2.814800e+02	1.779275e+04
cumulative_gas_co2	16359	0.35	2.587100e+03	1.234769e+04	0.00	3.24	5.206000e+01	4.577800e+02	2.452319e+05
cumulative_oil_co2	4665	0.81	3.296580e+03	2.164526e+04	0.00	3.92	3.916000e+01	3.727100e+02	5.926212e+05
cumulative_other_co2	23205	0.08	2.935900e+02	7.777100e+02	0.00	7.71	3.564000e+01	1.591900e+02	7.725990e+03
share_global_cumulative_cement_co2	12956	0.49	4.460000e+00	1.569000e+01	0.00	0.04	2.000000e-01	9.300000e-01	1.000000e+02
share_global_cumulative_coal_co2	8016	0.68	7.210000e+00	2.064000e+01	0.00	0.00	7.000000e-02	9.100000e-01	1.000000e+02
share_global_cumulative_flaring_co2	20822	0.17	5.620000e+00	1.502000e+01	0.00	0.06	5.400000e-01	3.570000e+00	1.000000e+02
share_global_cumulative_gas_co2	16359	0.35	5.240000e+00	1.893000e+01	0.00	0.01	1.100000e-01	8.200000e-01	1.000000e+02

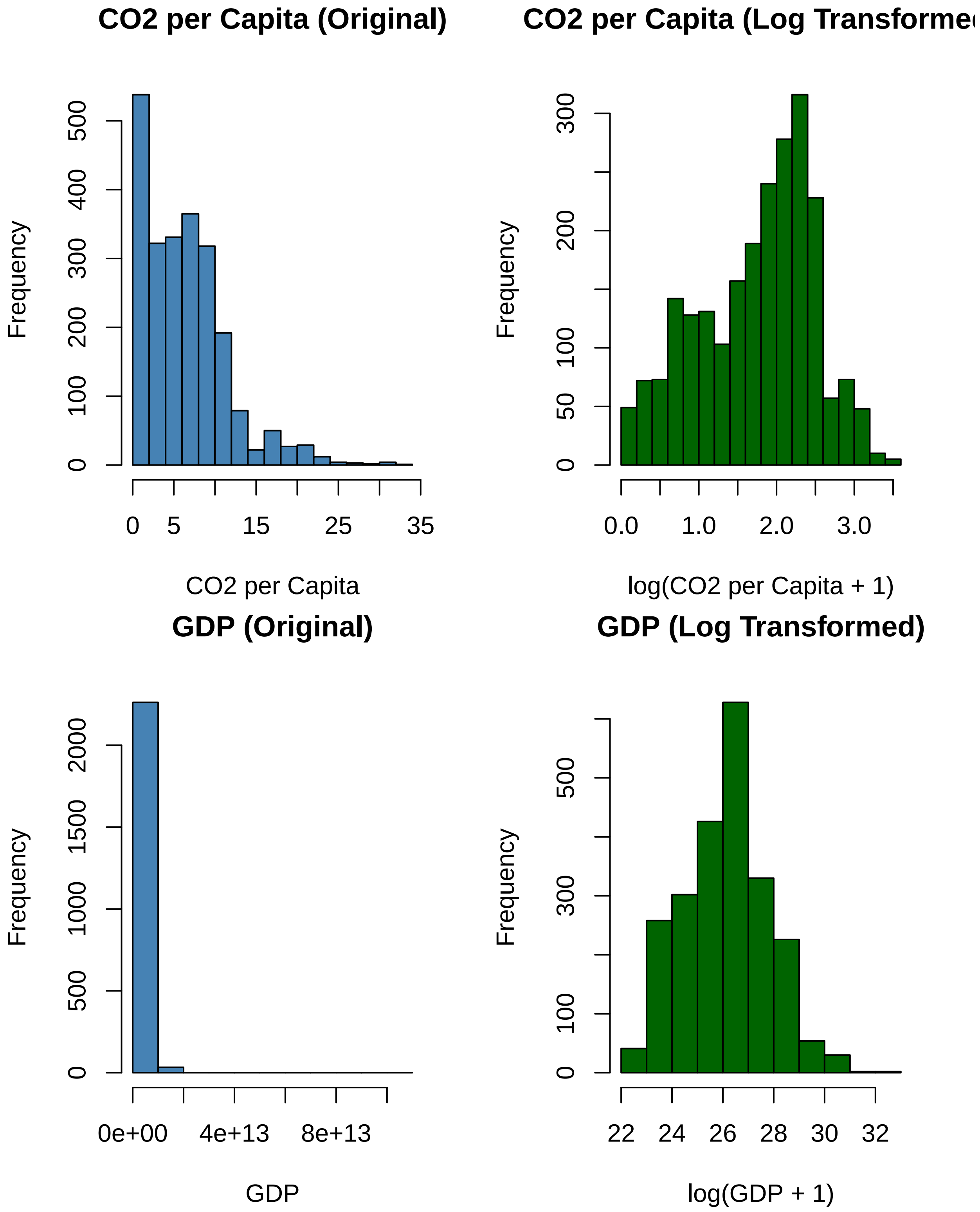
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
share_global_cumulative_oil_co2	4665	0.81	3.000000e+00	1.213000e+01	0.00	0.01	7.000000e-02	5.300000e-01	1.000000e+02
share_global_cumulative_other_co2	23205	0.08	1.340000e+00	2.702000e+01	0.00	0.19	8.400000e-01	7.980000e+00	1.000000e+02
total_ghg	19996	0.21	4.414800e+02	3.289040e+03	-85.62	8.16	3.709000e+01	1.322600e+02	4.935498e+04
ghg_per_capita	20049	0.20	8.100000e+00	9.390000e+00	-50.49	2.51	5.470000e+00	1.030000e+01	8.699000e+01
methane	19993	0.21	8.225000e+01	5.667000e+02	0.00	2.15	9.030000e+00	3.106000e+01	8.660010e+03
methane_per_capita	20047	0.20	2.050000e+00	3.530000e+00	0.00	0.72	1.110000e+00	1.690000e+00	3.981000e+01
nitrous_oxide	19993	0.21	2.931000e+01	1.992900e+02	0.00	0.54	3.590000e+00	1.057000e+01	3.054000e+03
nitrous_oxide_per_capita	20047	0.20	6.100000e-01	8.400000e-01	0.00	0.23	3.800000e-01	6.100000e-01	8.240000e+00
population	2326	0.91	7.072322e+07	3.795858e+08	1430.00	1291899.00	4.880320e+06	1.759622e+07	7.794799e+09
gdp	11666	0.54	2.877088e+11	2.180094e+12	55432000.00	9828867174.75	3.737019e+10	1.268944e+11	1.136302e+14
primary_energy_consumption	16514	0.34	1.569080e+03	9.366100e+03	0.00	7.00	6.140000e+01	3.528800e+02	1.621943e+05
energy_per_capita	16523	0.34	2.556849e+04	3.431996e+04	0.00	3270.37	1.370132e+04	3.549378e+04	3.475825e+05
energy_per_gdp	18401	0.27	1.850000e+00	1.580000e+00	0.05	0.86	1.410000e+00	2.350000e+00	1.349000e+01

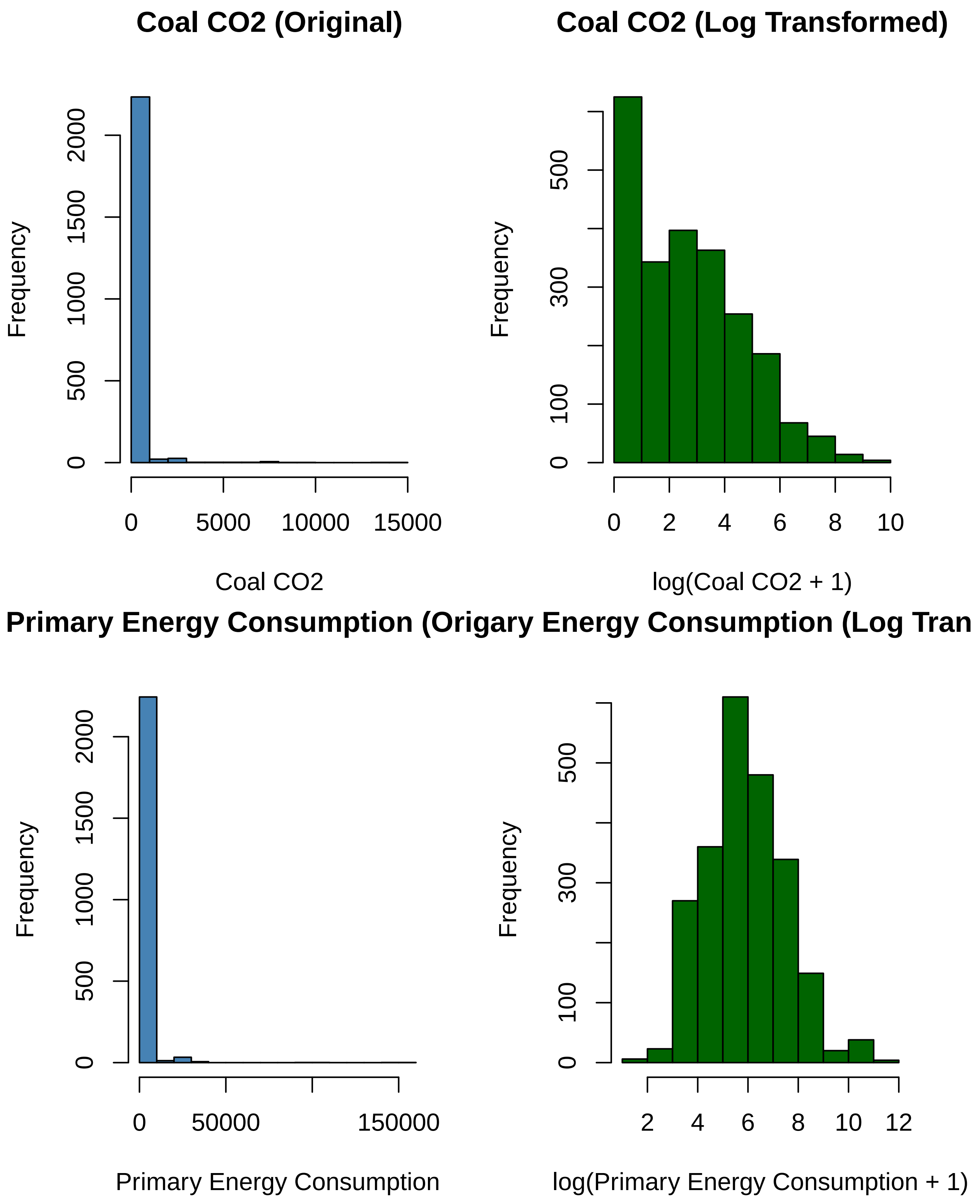
```
## # A tibble: 6 x 58
##   iso_code country     year   co2 consumption_co2 co2_growth_prct co2_growth_abs
##   <chr>    <chr>      <dbl> <dbl>           <dbl>           <dbl>          <dbl>
## 1 AFG      Afghanist~  1949 0.015              NA              NA         NA
## 2 AFG      Afghanist~  1950 0.084              NA             475          0.07
## 3 AFG      Afghanist~  1951 0.092              NA               8.7        0.007
## 4 AFG      Afghanist~  1952 0.092              NA               0          0
## 5 AFG      Afghanist~  1953 0.106              NA              16          0.015
## 6 AFG      Afghanist~  1954 0.106              NA               0          0
## # i 51 more variables: trade_co2 <dbl>, co2_per_capita <dbl>,
## #   consumption_co2_per_capita <dbl>, share_global_co2 <dbl>,
## #   cumulative_co2 <dbl>, share_global_cumulative_co2 <dbl>, co2_per_gdp <dbl>,
## #   consumption_co2_per_gdp <dbl>, co2_per_unit_energy <dbl>, coal_co2 <dbl>,
## #   cement_co2 <dbl>, flaring_co2 <dbl>, gas_co2 <dbl>, oil_co2 <dbl>,
## #   other_industry_co2 <dbl>, cement_co2_per_capita <dbl>,
## #   coal_co2_per_capita <dbl>, flaring_co2_per_capita <dbl>, ...
```

```
## [1] 2299   20
```



```
##             co2_per_capita                        gdp
##                   1.293015                  18.162317
##                 population                   coal_co2
##                  14.613508                  11.113842
##                    oil_co2                    gas_co2
##                  15.688054                  15.246990
## primary_energy_consumption          energy_per_capita
##                  14.731171                   1.780145
##             energy_per_gdp
##                   2.284947
```

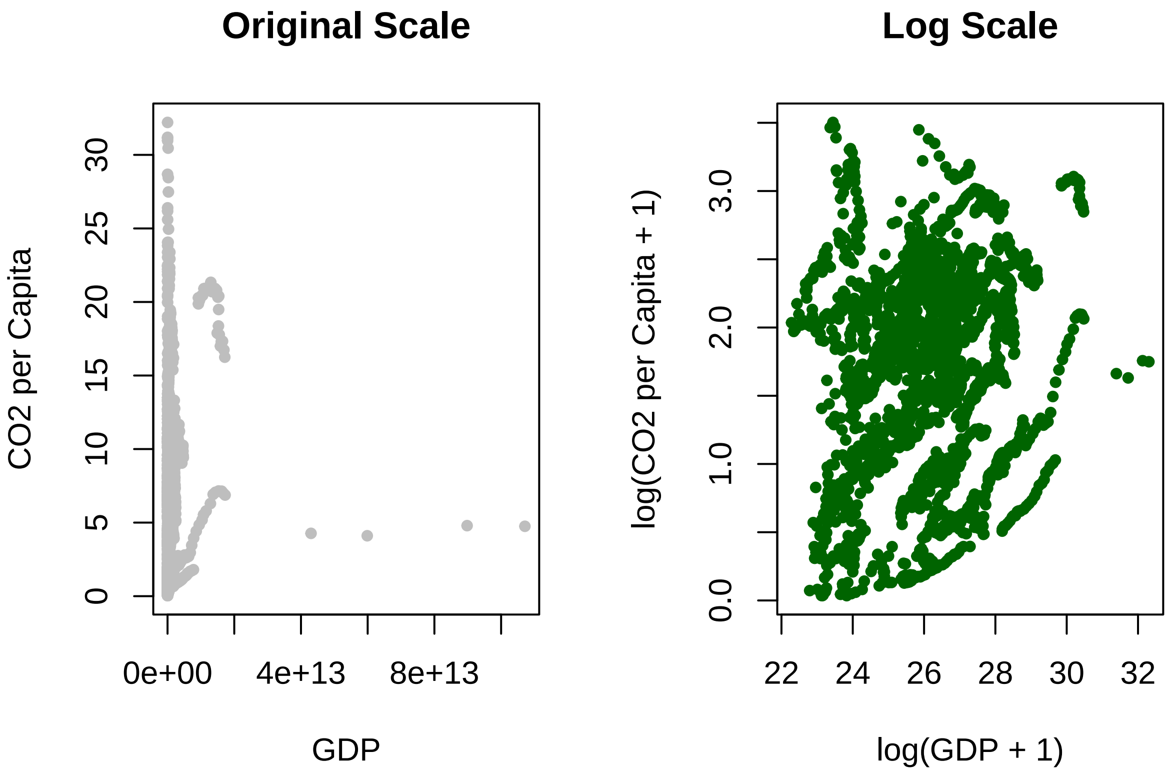

CO2 per Capita (Original)
Frequency
0
100
200
300
400
500
0
5
15
25
35
CO2 per Capita
CO2 per Capita (Log Transforme
Frequency
0
50
100
200
300
0.0
1.0
2.0
3.0
log(CO2 per Capita + 1)
GDP (Original)
Frequency
0
500
1000
1500
2000
0e+00
4e+13
8e+13
GDP
GDP (Log Transformed)
Frequency
0
100
300
500
22
24
26
28
30
32
log(GDP + 1)



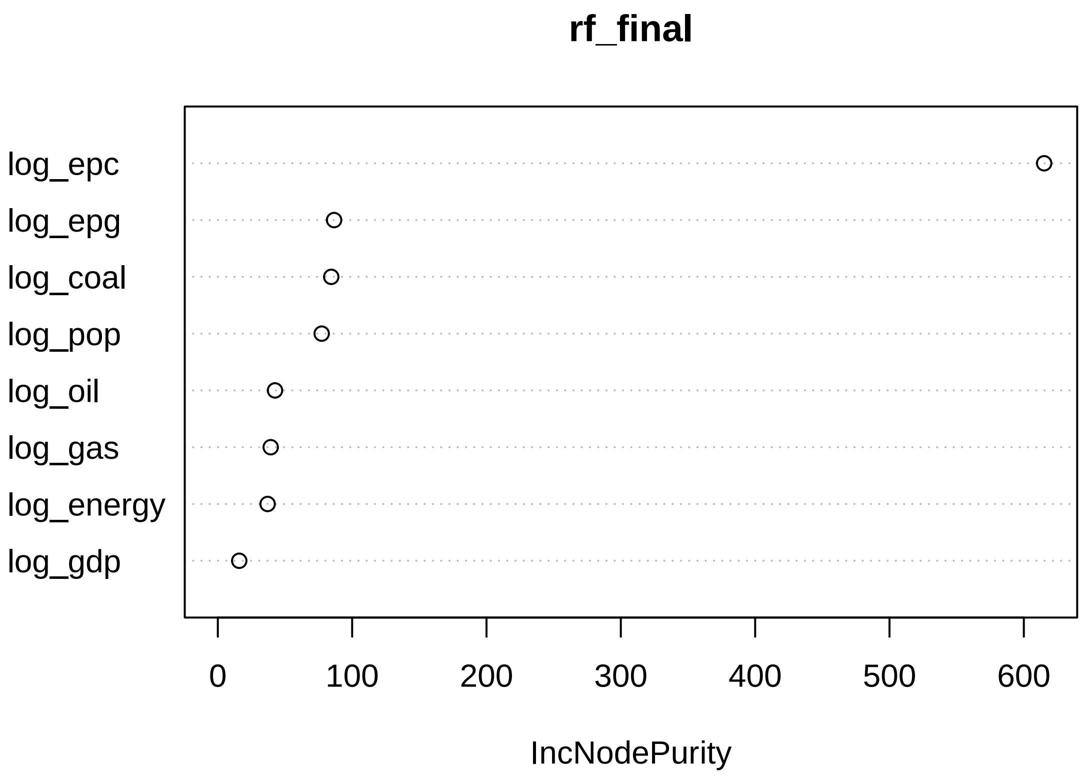
Coal CO2 (Original)

Coal CO2 (Log Transformed)

Primary Energy Consumption (Origary Energy Consumption (Log Tran



```
## [1] 0.06787627

## [1] 0.2222583

## [1] 0.2249576

## [1] 0.1703683

## [1] 0.1684915

##   mtry
## 2    3

## [1] 0.02816814

## [1] 0.06546055

## [1] 0.01772468

## [1] 0.03952533
```



```
## [1] 0.2222583
```

```
## [1] 0.2249576
```

```
## [1] 0.1703683
```

```
## [1] 0.1684915
```

```
## [1] 0.02816814
```

```
## [1] 0.06546055
```

```
## [1] 0.01772468
```

```
## [1] 0.03952533
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                               lambda.min
## (Intercept)                 1.140025e+00
## gdp                        -1.036215e-14
## population                 -1.852510e-10
## coal_co2                    6.051678e-05
## oil_co2                     8.612135e-06
## gas_co2                     1.022232e-04
## primary_energy_consumption -1.190284e-06
## energy_per_capita           4.556922e-06
## energy_per_gdp             -1.619635e-02
## log_gdp                    -3.131473e-02
## log_pop                    -1.518056e-01
## log_coal                    1.060061e-01
## log_oil                     9.242586e-02
## log_gas                     6.627011e-02
## log_energy                 -1.033295e-02
## log_epc                          3.138034e-01
## log_epg                         -1.049698e-03
```