## Markdown Explanation of Our Process

We chose our project because renewable energy adoption is becoming increasingly important worldwide. Governments significantly influence this adoption through policy effectiveness and transparency, captured by the “Regulatory Quality” index. To explore how regulatory quality impacts renewable energy use, we selected several relevant variables:

- Regulatory quality index
- $CO_2$ emissions (% change from 1990)
- GDP growth (% annual)
- Fossil fuel energy consumption (% of total energy)
- Foreign direct investment (FDI)
- Control of corruption
- Energy imports (% net energy use)
- Energy efficiency (energy use per GDP)
- Renewable freshwater resources

We gathered datasets from the World Bank, focusing on reliable data availability across countries and multiple years.

## Initial Data Exploration

We first reshaped our data from wide to long format to standardize and merge them. An initial look revealed significant missing data, particularly in the regulatory environment rating (~71%) and energy imports (~43%). To handle this, we chose the top 50 countries with the least missing values, ensuring a balanced dataset.

A correlation matrix was generated, showing moderate correlations, but none of them showing high enough correlation that required excluding variables. Preliminary visualizations (scatter plots, trend lines) suggested a clear positive relationship between regulatory quality and renewable energy usage.

## Data Cleaning and Encoding

We created a cleaned dataset (df_cleaned) by converting all non-numeric variables (except country names) to numeric. After assessing missing data, we focused our analysis on the top 50 countries with the most complete data. We further categorized these countries by income groups (Low, Lower-middle, Upper-middle, and High income) using World Bank classifications.

For modeling, we encoded income groups using an ordinal encoding scheme, capturing the economic development gradient. Numerical variables were standardized with StandardScaler to prepare for linear regression models.

## Modeling, Model Validation & Selection

We split our data into training and testing sets (80%-20%). Two modeling approaches were used:

1. Linear Models:

- Linear Regression: Provided a baseline. The initial model showed reasonable performance.

- Elastic Net Regression: We tuned hyperparameters (alpha, l1_ratio) via grid search to balance bias and variance effectively.

2. Random Forest Regression:

To capture potential nonlinear relationships, we employed a Random Forest model. We optimized hyperparameters (such as tree depth and sample splits) using a grid search approach.

**Keynote**: For linear regression and elastic net, we dropped missing values to ensure precise results. For the random forest, we retained more data by not dropping missing values post-grouping to use its flexibility in handling incomplete data.

**Model Assessment**

We assessed our models based on Mean Squared Error (MSE) and the $R^2$ (coefficient of determination), where a higher $R^2$ (closer to 1) and lower MSE indicate better predictive performance:

Linear Regression:

- MSE: 82.56
- $R^2$: 0.8664

The linear regression provided a solid baseline, explaining approximately 86.6% of the variability in renewable energy usage.

Elastic Net Regression:

- Best Parameters: $\alpha = 0.01$, L1 ratio = 0.9
- MSE: 82.45
- $R^2$: 0.8666

Elastic Net slightly improved on the linear regression, offering very similar performance but with better generalization due to regularization.

Random Forest Regression:

- Best Parameters: 200 estimators, unlimited max depth, min samples leaf = 1, min samples split = 2
- MSE: 13.55
- $R^2$: 0.9781

The Random Forest significantly outperformed linear models, explaining about 97.8% of variance in renewable energy consumption. Its lower MSE also suggests stronger predictive accuracy, highlighting that renewable energy adoption involves complex, non-linear relationships not captured fully by linear models.

Overall, the Random Forest regression was our most effective model, clearly illustrating the importance of non-linear modeling for this analysis.

In conclusion, regulatory quality strongly correlates with renewable energy adoption, especially when considering non-linear effects and differences across income groups. This analysis provides a solid foundation for further, more detailed policy-oriented studies.